FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Adaptive phonetic segmentation in dysphonic voice

João Filipe Torres Costa

MASTER IN ELECTRICAL AND COMPUTERS ENGINEERING

Supervisor: Prof. Aníbal João de Sousa Ferreira

Co-Supervisor: Engº João Miguel Pinto Pereira da Silva

12 March 2021

Resumo

A fala sussurrada é normalmente usada em muitas interacções, designadamente em contextos onde os oradores tentam ser discretos ou onde o silêncio é requerido. Contudo, para aqueles que sofrem de disfonia ou afonia, o sussurro é a sua única forma de comunicação disponível, o que pode causar problemas a nível social e pessoal. Actualmente, nenhuma solução do mercado para conversão de fala sussurrada em fala sonora é satisfatória devido à sua falta de naturalidade e capacidade de resposta da fala ou à sua implementação ser intrusiva para o paciente. Esta tese, integrada no projecto DyNaVoiceR (PTDC/EMD-EMD/29308/2017), visa resolver uma parte do problema através da criação de um algoritmo capaz de realizar a segmentação da fala sussurrada em tempo real. Além disso, este é um projecto pioneiro em Portugal, onde não existe nenhuma solução implementada para pacientes disfónicos.

Para este fim, propomos construir, em primeira instância, um conjunto de regras eficientes e simples para a deteção de consoantes oclusivas não vozeadas, das fricativas e sibilantes, a fim de serem conhecidas as características associadas a estes tipos de fonemas. Na segunda instância, utilizando os conhecimentos adquiridos, é construído um modelo estocástico, os *Hidden Markov Models*, com os estados do silêncio e vogais acrescentados, com *features* previamente aprendidas, sendo estas complementadas pelos coeficientes de Mel.

Utilizando o conjunto de regras simples, obtivemos taxas de acerto de 94%, 77% e 85% para as consoantes oclusivas não vozeadas, fricativas e sibilantes, respectivamente.

Para a segunda abordagem, os HMMs obtiveram uma taxa de acerto de 85%, o que é substancial devido a algumas características comuns de alguns fonemas presentes. Note-se também que a utilização de HMMs permite a execução em tempo real, comprovada pelo tempo de execução de 18 ms por amostra, que é inferior à janela em análise, de 23 ms, à taxa de amostragem de 22050 Hz.

Estes resultados provam que com a utilização de regras simples pode-se alcançar grandes resultados e será a base de estudos futuros para os restantes fonemas que compõem a língua portuguesa.

Abstract

Whispered speech is commonly used in many interactions, namely in the context where the speakers try to be discreet or where silence is mandatory. However, for those who suffer dysphonia or aphonia, whispered speech is their only form of communication available, which can cause problems at a social and personal level. Currently, all market solutions implementing whispered-speech to voiced-speech conversion are not satisfactory due to their lack of speech naturalness and responsiveness or due to the fact that their implementation is intrusive to the patient. This thesis, integrated into the DyNaVoiceR (PTDC/EMD-EMD/29308/2017) project, which is a project that aims to convert whispered speech to natural speech in real time, intends to create an algorithm capable of performing segmentation for whispered speech in real-time. DyNaVoiceR, is a pioneer project in Portugal, as no solution available has been implemented for dysphonic patients.

For this purpose, we propose to build in a first instance an efficient and simple set of rules for the unvoiced stop consonants, fricative and sibilant phonemes in order to understand the characteristics associated with these phonemes. In a second instance, using the knowledge acquired, a stochastic model, the Hidden Markov Models, is built, with silence and vowels added, with features previously learned supplemented by the Mel-frequency cepstrum coefficients.

Using the simple ruleset, we obtained score rates of 94%, 77% and 85% for the unvoiced stop consonants, fricative and sibilant phonemes, respectively.

For the second approach, the HMMs managed to have a score rate of 85%, which is substantial due to some shared features some phonemes present. It is also noted that using HMMs allow execution in real-time, which is proven by the execution time of 18 ms per frame, which is inferior to the window in analysis, of 23 ms, at the sample rate of 22050 Hz.

These results prove that using simple features allows achieving useful results and will be the basis of future studies for the remaining phonemes that make up the Portuguese language.

Agradecimentos

Aos meus orientadores, Doutor Aníbal Ferreira e Eng. João Silva gostaria de agradecer por me integrarem no projeto e orientarem a minha tese. Agradeço todo o aconselhamento,a orientação, sugestões, motivação e camaradagem.

Aos meus companheiros neste percurso académico, a todos que já o completaram e os que ainda o caminham com o fim em vista, o meu muito obrigado. Sem vocês esta viagem teria sido muito mais custosa e sem o divertimento e companheirismo habitual.

À minha namorada, o meu muito obrigado por me fazeres acreditar que eu tinha o necessário para ultrapassar as adversidades que me foram aparecendo.

À minha irmã, o meu muito obrigado pelo apoio e preocupação constante ao longo destes anos.

Por fim, aos meus pais, obrigado por acreditarem em mim e ajudarem-me a tornar o homem que sou hoje.

João Costa

"The most valuable of all talents is
that of never using two words when one will do."

Thomas Jefferson

Contents

List of Figures

List of Tables

Abbreviations

DFT	Discrete Fourier Transform
f_o	Fundamental Frequency
HMM	Hidden Markov Models
IPA	International Phonetic Alphabet
LBDP	Level Building Dynamic Programming
LDA	Latent Dirichlet Allocation
I MAP	*maximum a posteriori* Probability
MFCC	Mel-frequency Cepstrum Coefficients
ODFT	Odd Discrete Fourier Transform
PMTK	Probabilistic Modeling Toolkit for Matlab/Octave
SFM	Speactral Flatness Measure
SPF	Female Speaker
SPM	Male Speaker
STM	Spectral Transition Measure
VCP	Vowel-consonant-pause
VOT	Voice Onset Time

Chapter 1

Introduction

This introductory chapter presents the dissertation context, its motivation and goals, as well as the document structure.

1.1 Context

Around 80% of all human voice sounds, using as a reference the sound diversity that exists in regular speech, are produced by vibrating the vocal folds. These sounds are called sonorous or voiced. The other 20% include non-sonorous or unvoiced sounds, the only ones found in whispered speech. These sounds result mostly from turbulent noise generated at the glottis when the vocal folds do not vibrate. Also, they can be produced by merely moving some parts of the mouth, like the teeth. Similarly to those, there are also occlusive sounds that are formed by accumulating air with the mouth closed and then suddenly releasing it [3].

Unfortunately, some individuals are restricted to only using the whispered speech due to some aphonia or dysphonia, impacting their lives personally and professionally. These include Paradoxical Vocal Fold Movement [4], Vocal Fold Nodules and Polyps [5], Spasmodic Dysphonia [6] and Vocal Fold Paralysis [7].

Current solutions to this type of conditions include the Electrolarynx (EL) [8], whose speech is notorious of the sound quality being monotonic and robotic with the lack of pitch control and the presence of the radiated noise. Other solutions that rely on technology include the use of silent speech interfaces, as well as text-to-speech (TTS) [9] applications. The main problems with the current solutions that the DyNaVoiceR research project will attempt to solve is the lack of real-time operation as well as the lack of identity on the replacement voice. This thesis will contribute to this solution by applying correct segmentation of the whispered speech.

1.2 Motivation

The impossibility of effectively communicating is an extreme disadvantage to any individual socially and professionally. Mitigating this problem in a non-evasive and practical way, conserving

the voice identity and operating in real-time, goes a long way to improve the quality of life of the subjects in these conditions. This research work aims to create non-invasive and quality solutions that solve most of these problems.

1.3 Objectives

This thesis proposes to identify a set of features and recognisable patterns found in whispered speech words. In order to carry out the task above, words' waveforms are split into windows with a size of 1024 samples and an overlap-add of 50%. The overlap is employed to increase the robustness of the analysis, by not discarding the signal information on the edges of the window. Then, applying the Odd Discrete Fourier Transform (ODFT), data like the signal energy and phase is gathered. This information is the baseline for identifying the features and patterns on the different words and sentences in the DyNaVoiceR database and applying the segmentation to the whispered speech based on that. The segmentation will be conducted using Hidden Markov Models and simple rules to increase the set of solutions to the problem risen in this dissertation.

Having in mind a real-time application, these techniques must be simple and computationally efficient.

1.4 Document Structure

The dissertation is divided in the following manner:

- **Chapter 1**: Describes the dissertation context, motivation and goals;
- **Chapter 2** : Presents an introductory analysis of the topics achieved in this dissertation, as well as the current state of art regarding those topics;
- **Chapter 3**: Presents the ruleset used and results for the segmentation of unvoiced stop consonants;
- **Chapter 4**: Presents the ruleset used and results for the segmentation of fricatives and sibilants;
- **Chapter 5**: Introduces HMMs and the approach used to detect 6 states representing Portuguese Phonemes as well the results obtained;
- **Chapter 6**: Concludes the dissertation and proposes the future work.

Chapter 2

State Of Art

This chapter introduces the topics this dissertation addresses. It is divided into two main categories, the production of speech and its associated systems and the phonetic segmentation works done in the area of expertise. Each area has particularities that will help to get the necessary information to develop a set of rules or decisions required to build the algorithms developed in this dissertation.

2.1 Contextualisation

This section presents the topics that contextualise the production of speech. From the manner speech is produced, how it is perceived, models representing it, and the Portuguese language's phonetics. This contextualisation gives an overview of the thematic and is the following chapters' foundation.

2.1.1 Speech Production

Speech communication plays the single most crucial role in human nature. Without speech, the social aspect of human existence would be diminished. Consequently, the success of human evolution can be explained in part by it due to the cooperation intrinsic to our species in which speech plays an important role.

The human vocal apparatus that is represented in Figure 2.1, is the system responsible for human speech production. The apparatus includes a source of air flow (the lungs), components that vibrate (the vocal folds in the larynx) and resonant chambers (the pharynx, the mouth, and the nasal cavities).

In speech production [10], the lungs (element of the subglottic system) provide the air necessary, working as a "generator" to provide enough volume and pressure of airstream to vibrate the vocal folds, producing a periodic pattern, especially in vowels and voiced consonants. This pattern is known as pitch and is defined by the vocal folds length, despite speakers having the ability to introduce some variability. The pitch is the auditory human perception of the fundamental frequency (f_o). Afterwards, the vocal folds' acoustic signal is modulated by the vocal tract's elements, creating different phonemes.

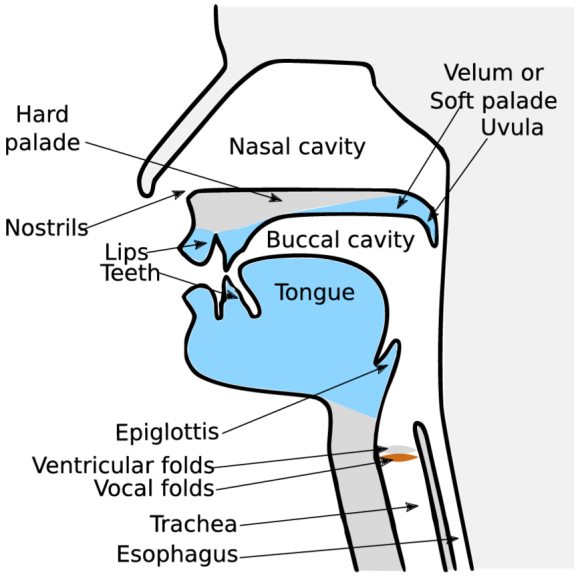

Figure 2.1: Human Vocal Apparatus (adaptation from [1])

However, in the whispered speech scenario, the vocal folds are abducted, so they do not vibrate. Only the supralaryngeal articulation remains the same. Therefore, whisper speech shares some similarities with voiced speech but contains significantly fewer energy and the characteristic spectral range is not given due to the absence of tone [11]. Due to these properties, implementing speech recognition in whisper proves to be more challenging than in the voiced counterpart.

2.1.2 Psychoacoustics

Psychoacoustics is a branch of psychophysics that studies the perception of acoustic signals. It builds quantitative models that relate psychological responses with the physical characteristics of signals [12] [13] [14]. The acoustic signal amplitude relates with human ear acoustic perception by following a logarithmic relation, thus the use of the deciBel scale (dB) to represent the magnitude of acoustic signals. The f_o is interpreted in audition as the tone and, due to human ear properties, follows a non-linear relation with frequency. This is where quantitative models are important, and one of the scales built around this difference in human ear perception and the linear frequency is the Mel scale. This scale is a perceptual scale of pitches calculated using the formula 2.1, given by Gunnar Fant [15], that gives a rough approximation of how the human ear "hears" specific frequencies. For example, the human ear can notice differences of 1 Hz in low frequency sounds, but the same does not happen in higher frequencies, where higher frequency steps are needed for human perception.

$$m = \frac{1000}{log2} log\left(1 + \frac{f}{1000}\right) \tag{2.1}$$

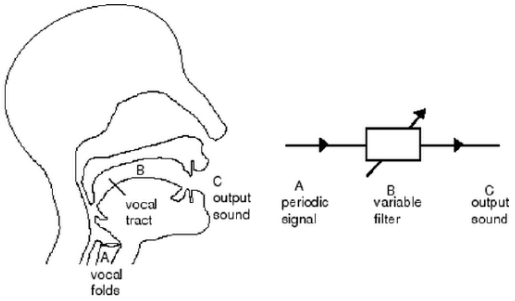

Figure 2.2: Source-Filter Model (adaptation from [2])

2.1.3 Source–Filter Model

The source-filter model, introduced by Gunnar Fant [16], describes the human vocal apparatus as a combination of sound sources and acoustic filters. The sound sources are divided into periodic, corresponding to the signal components produced by the vocal folds, and are responsible for the voicing and non-periodic, corresponding to the non-voiced speech types.

The filter represents all the supraglottic system and acts as an acoustic filter for every human's sound, defining the sounds' spectral envelope produced during the speech. This model acts independently between the source and the filter.

Despite being an approximation, this simple model represents speech production conveniently, allowing modelling speech in both voiced and whispered scenarios. This paradigm serves as the base to cepstral analysis [17], which is widely used in speech segmentation, that also proved to be useful in our case, as shown in the following chapters. Figure 2.2 illustrates the source-filter model.

2.1.4 Phonetics

Phonology studies languages' sound systems based on the speakers intuitive and mental knowledge of their language. It is the theoretic study of any language and, in this case, the Portuguese language. Phonetics focus on the study of speech sounds, production, and perception. It is divided into articulatory phonetics (the study of the movement of articulators for the production of speech sounds), acoustic phonetics (the study of speech sounds' physical properties) and perceptive phonetics (the study of how the speech sounds are heard and interpreted). In this section, we will focus on the study of articulatory phonetics in the Portuguese language [18]. It subdivides into two significant parts:

- **Vowels:** Vowels are produced without any significant constrictions to the airflow of the oral tract and subdivided by the tongue's height (Close, close-mid, open-mid and open) and articulation point (Front, Central, Back). In the case of Portuguese language, there are two types of vowels, oral vowels (Table 2.1) and nasal vowels (Table 2.2), where the airflow on its way out is also expelled by the nasal tract.

	Front	Central	Back
Close	i	ɨ	u
Close-mid	e		o
Open-mid	ɛ	ɐ	ɔ
Open		a	

Table 2.1: Oral Vowels

	Front	Central	Back
Close	ĩ		ũ
Close-mid	ẽ		õ
Open-mid		ɐ̃	

Table 2.2: Nasal Vowels

- **Consonants:** Contrary to the vowels, consonants are produced with substantial constrictions to the oral tract's airflow due to articulators movement. The articulators can narrow or prevent the oral tract's airflow completely, creating noise-like sounds. The parameters that distinguish Portuguese consonants' articulatory production are:

  - articulation point (labial, dental/alveolar, dorsal, plain and labialized);
  - articulation mode (plosive, fricative, approximant and rhotic);
  - soft palate position (obstructing and sounding);
  - vocal folds state (voiced, voiceless).

  The articulation mode classification is based on the oral tract perturbation to the airflow and will prove to be the basis of this dissertation, as the following chapters show. Hence, it is of our interest to distinguish the characteristics of each articulation mode subcategories for consonants:

  - Plosives or stop consonants: Airflow is completely constricted;
  - Fricatives: Partial airflow constriction, creating some noise;
  - Approximants: Central constriction to the airflow, forcing the airflow to move through the tongue' sides;
  - Rhotic: Partial airflow constriction that induce tongue vibration.

  Table 2.3 shows all the Portuguese consonants grouped:

		Labial	Dental/ Alveolar	Dorsal	
				plain	labialized
Nasal		m	n	ɲ	
Plosive	voiceless	p	t	k	(*k***w**)
	voiced	b	d	g	(*g***w**)
Fricative	voiceless	f	s	ʃ	
	voiced	v	z	ʒ	
Approximant	semivowel			j	w
	lateral		l	ʎ	
Rhotic	trill/fricative		ɾ		
	flap		ʀ		

Table 2.3: Portuguese consonant phonemes

2.2 Literature Review

Nowadays, there are several techniques in the area of speech recognition and speech segmentation. This section will discuss the relevant bibliography, whose theme overlaps with the one of this dissertation. This work will serve as a basis for the dissertation and provides new techniques to be explored in later work stages.

2.2.1 Detection of Stop Consonants

Firstly, the work carried out in [19] approaches the detection of stop consonants by using the burst of the signal, formant transitions and a combination of the two as features. The results were obtained by training on citation-form data and continuous speech data. The results show that there is more information for the place distinction in the burst than in formant transitions. When the parameters are combined into a single model, classification scores are improved for the citation-form data but not for the continuous speech data, reaching a hit rate of 90%. In [20] a general framework is described to parameterize the speech waveform in terms of linguistic features using HMMs with Gaussian models. According to the authors, this approach gives good results, but they don't quantify the result. In [21], the voice onset time (VOT) and the duration of the burst of frication noise at the release of a plosive consonant were measured from spectrograms of consonant clusters. Although this may be a good feature to detect stop consonants, it cannot be used in this thesis's due to current frames depending on future ones, which cannot be used in a real-time environment. Finally, [22] uses articulator-free features, such as energy abruptness of five frequency bands and two levels of temporal resolution, segmental duration, and broad phonetic class constraints, and articulatory constraints. It is noted that the time features mostly produce results in 20ms of the transcription, while not being real-time, has some potential to be used as such.

2.2.2 Detection of Fricatives

According to [23], fricatives' prime acoustic characteristics are the concentration of spectral energy above 3 kHz and having a noisy feature. Their work explores using the S-transform for detecting fricatives from continuous speech due to its properties being tailored for localizing high-frequency events. No mention of performance is made. In [24], two fricatives and affricates algorithms are presented. One is based on the cepstrogram-matching approach and the other on an LDA classifier with a feature vector based on the audio signal's temporal, spectral and textural features. Once more, temporal features are unusable in our use-case. Despite that, the algorithm had an identification rate of over 90% and a specificity of over 85%. Finally, in [25] it is also referred, as mentioned in [23], that fricatives have energy across the noise spectrum. Moreover, it is mentioned that the formant transitions play a role in identifying some fricatives, but combining both features creates conflicts in the final results.

2.2.3 Automatic Phoneme Segmentation

After searching about the specific phonemes to be studied in this dissertation, a general survey about the methods used for general automatic phoneme recognition was done. In [26] it is used the Spectral Transition Measure (STM) and Level Building Dynamic Programming (LBDP). Both algorithms use spectral variation as the base to find the boundaries of a phoneme. Both algorithms have their performance analysed after extracting 12 MFCCs with a performance of 77.8% for STM and 79.3% for LBDP. In [27], a simple VCP (vowel-consonant-pause) segmentation is performed for the Indian language. The method applies velocity and acceleration parameters of rate-of-change dynamics on formants of speech with performance variable between 60% and 90%. In [28], a language-independent solution is based on HMMs with 9 phonetic classes (9 states). For this solution, it was obtained an 80% average recognition. Finally, in [29], a similar solution to [28] is applied but for spectrogram images with good results but a high number of false positives, deeming the results unusable.

2.3 Summary

The contextualisation allowed obtaining a higher acquaintance of the studied theme, especially the different phonemes of the Portuguese language, especially the consonants, due to the future need of getting a grip of them in the following chapters. The literature review allowed knowing a set of methods and models that could prove helpful by giving us tools to branch the segmentation techniques used.

Chapter 3 and 4 will present the ruleset used to detect stop consonants, fricatives, and sibilants. Alternatively, a phonetic segmentation based on HMMs will be proposed on chapter 5.

Chapter 3

Stop Consonants Detection

This chapter describes a simple ruleset created based on signal observations present in the stop consonant DyNaVoiceR project database words.

3.1 Database

For the purpose of the DyNaVoiceR project, a database of European Portuguese recordings has been created. It includes recordings of 9 vowels, 4 single sibilants, 27 words, 6 sentences and a full text, all produced by 20 speakers (10 female and 10 male) in both voiced and whispered speech. The audio files were recorded using the audio format .wav at a sampling rate of 22050 Hz and are supplemented with files including manual annotations, allowing to locate and identify each phoneme present in each word of the recordings.

An empirical examination of all stop consonants present in the DyNaVoiceR database was conducted to get some awareness of their signal characteristics. Table 3.1 shows all 6 words containing unvoiced stop consonants in the database, their phonetic sequence, and the corresponding IPA symbol [30].

Furthermore, examples of the signal and spectrogram of each unvoiced stop phoneme, in both voiced and whispered speech regarding speaker SPM01 of the DyNaVoiceR database, are shown from Figures 3.1 to 3.3. In all portions of the figures corresponding to the whispered speech (right

Words	Phonetic Sequence	IPA Symbol
<nu/**c**/a>	VCV	k
<lu/**p**/a>	VCV	p
<ri/**p**/a>	VCV	p
</**p**/i/**c**/a>	CVCV	p,k
<lu/**t**/a>	VCV	t
<ri/**t**/a>	VCV	t

Table 3.1: Database Stop Consonants

side), it can be observed that there is a lack of harmonic component in the vowels' signal when compared to the voiced counterparts. However, the same does not occur in the stop consonants, suggesting that the same analysis used in voiced speech can be used in whispered speech.

After analysing all the data gathered, a few preliminary conclusions were made:

- There is a characteristic silence before each stop consonant.
- There is a burst of energy after the silence.
- The burst behaves like an impulse-like signal, which means that it can mathematically be replicated.

These conclusions will be the foundation of the rules built in the next section.

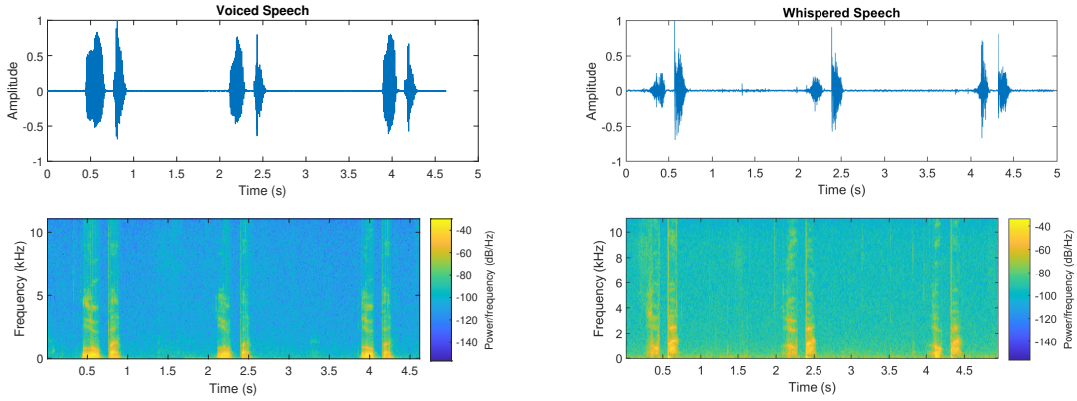

Figure 3.1: Recording of <lu/p/a> by SPM01

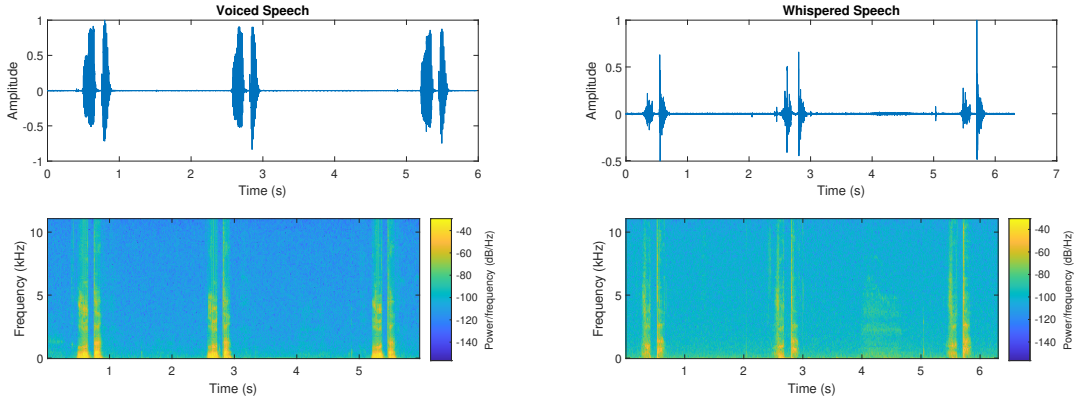

Figure 3.2: Recording of <lu/t/a> by SPM01

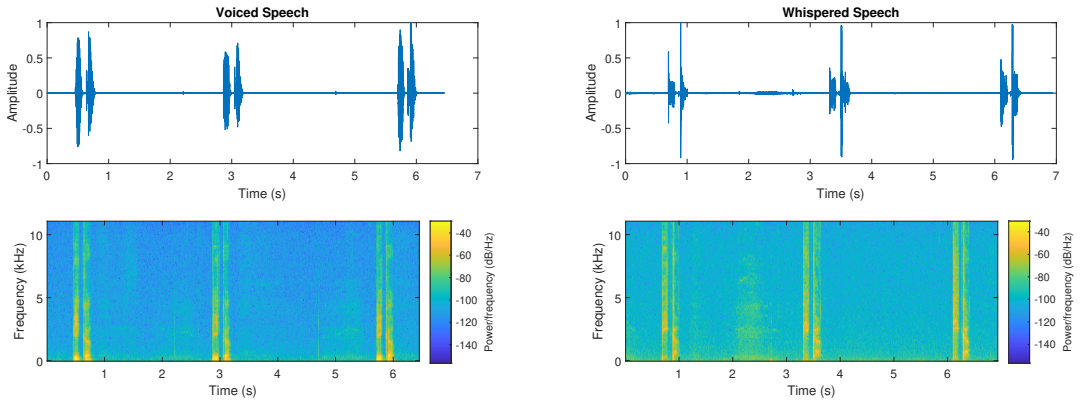

Figure 3.3: Recording of </p/i/c/a> by SPM01

3.2 Methodology

In this section, the methodology used for the detection of unvoiced stop consonants is presented. Firstly, the detection algorithm is shown through a block diagram, then each individual detection rule and their combination is presented afterwards.

3.2.1 Block Diagram

Figure 3.4 represents the unvoiced stop consonants detection algorithm's block diagram.

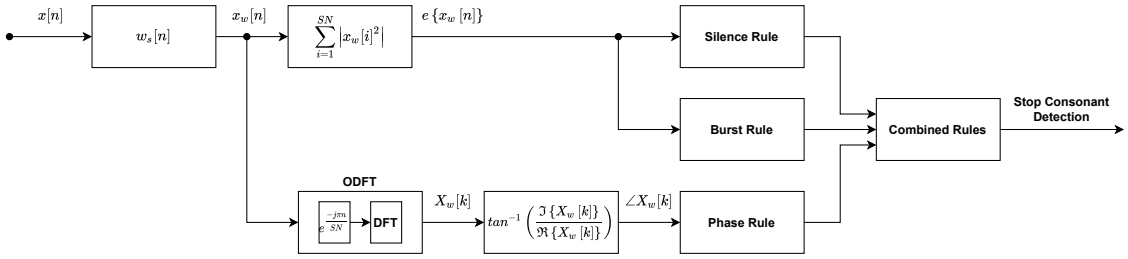

Figure 3.4: Unvoiced stop consonants detection algorithm's block diagram

The signal x[n] represents the input signal, sampled at 22050 Hz of frequency, which is subject to a sliding window analysis based on sinusoidal windows $w_s[n]$ with 256 samples and 50% signal overlap. We obtain the signal $x_W[n]$ for each window, which is then used to calculate the energy in the temporal domain and the ODFT. The latter is used to calculate the phase $\angle X_w[k]$. These characteristics will be used in the following subsection 3.2.2.

Figure 3.5 shows a graphical environment example of the micro-analysis, using 4 short windows of 256 sample with 50% overlap.

The example shown corresponds to the detection of the unvoiced stop consonant p for the word **p**ica, recorded in whispered mode. In the subplots presented, we represent the signal waveform after applying the sinusoidal window $x_W[n]$ and its corresponding frame energy in the time domain.

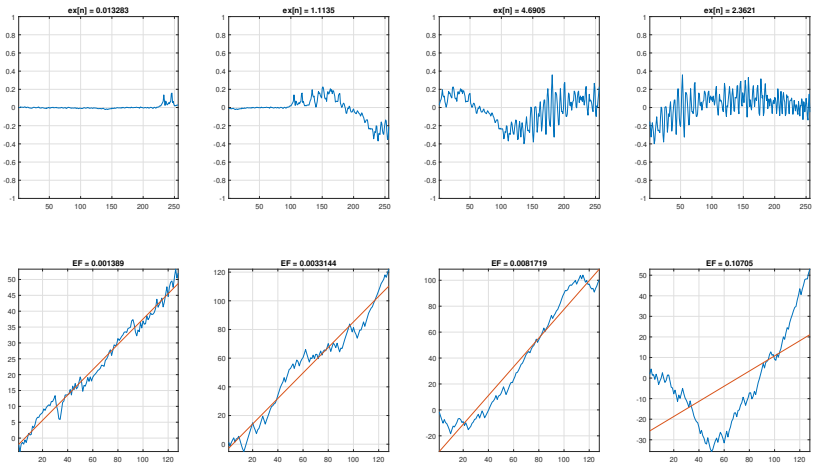

Figure 3.5: 4 windows of pre-processing visualisation

In addition, the subplots show the phase structure $\angle X_w[k]$ (in blue) and its linear model (in orange). The phase error will be explained in the next subsection. This representation proves to be useful because it allows analysing in detail each frame, allowing a better grasp for detecting errors and optimising the detection rules.

3.2.2 Signal Characteristics and Rules

This subsection presents the subset of rules developed to detect 3 characteristics of the unvoiced whispered stop consonants: the presence of silence ahead of the stop consonant, the occurrence of burst in the signal and the identification of a pulse-like signal employing phase modelling. Once the signal is processed, having a sliding window analysis with temporal overlap, the developed rules' output is identified separately for even windows (in red) and odd windows (in blue) in this subsection's figures. The grey areas indicate the presence of an unvoiced stop consonant as indicated by the manual annotation. Once more, the example given is for the word **pic**a.

- **Silence:** The presence of silence ahead of the signal burst characterises the unvoiced stop consonants. From the analysis of the DyNaVoiceR project database, the period of this silence is always greater than 23 ms. This period corresponds to 3 windows of 256 samples with 50% overlap for the sampling rate used of 22050 Hz. Equation (3.1) represents those characteristics for an energy threshold ($tr1$) that represents a low energy frame:

$$e\{x[k-3]\} \ \& \ e\{x[k-2]\} \ \& \ e\{x[k-1]\} < tr1 \tag{3.1}$$

  Where:

  - e : window energy;
  - x : audio signal in time domain;

– $tr1$: threshold 1;

Figure 3.6 illustrates the detection of the rule described.

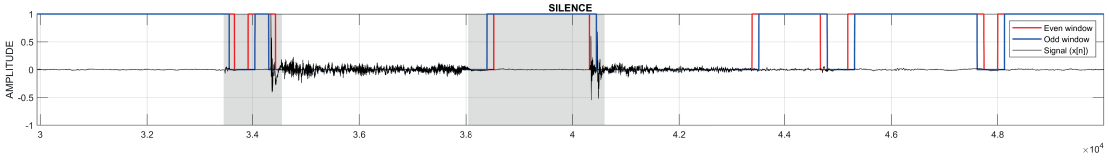

Figure 3.6: Silence Rule Detection

- **Phase:** Unvoiced stop consonants are characterised by being impulse-like signals, meaning that a complex exponential can model them. The unwrapped phase of a complex exponential (e^{jwn_0}) corresponds to a straight line with a positive or negative slope [31]. Depending on n_0 being positive or negative the phase model's detection rule verifies if the deviation between the phase structure obtained from the ODFT of the window and its theoretical linear model is inferior to a threshold ($tr2$). This deviation, or phase error (EF), is given by:

$$EF = \frac{\frac{\sum(model - \angle X[k])^2}{SN/2}}{max(abs(\angle X[k]))^2} \tag{3.2}$$

Where:

– EF : phase error;
– $model$: phase model;
– SN : short window size;
– X : audio signal in frequency domain;

The phase condition is given by:

$$EF < tr2 \;\&\; max(abs(\angle X[k])) > 1.0 \tag{3.3}$$

Where:

– EF : phase error;
– $tr2$: threshold 2;
– X : audio signal in frequency domain;

Figure 3.7 illustrates the detection rule for the phase model described.

- **Burst:** Unvoiced stop consonants are usually characterised by a sudden increase of the signal enery (burst). The equation that verifies this condition by comparing the energy of a given window k and the previous $(k-1)$ and verifying a minimal energy level $(tr3)$ for window k, is given by:

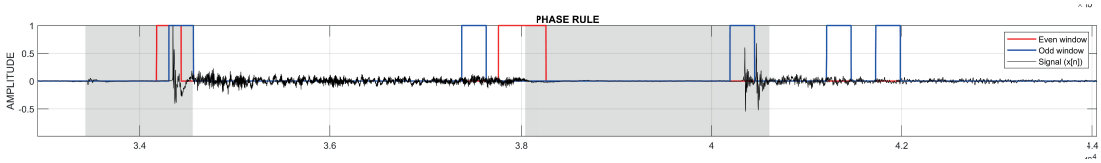

Figure 3.7: Phase Model Rule Detection

$$e\{x[k]\} > e\{x[k-1]\} \;\&\; e\{x[k]\} > tr3 \tag{3.4}$$

Where:

- e : window energy;
- x : audio signal in time domain;
- $tr3$: threshold 3;

Figure 3.8 illustrates the detection rule for the signal burst.

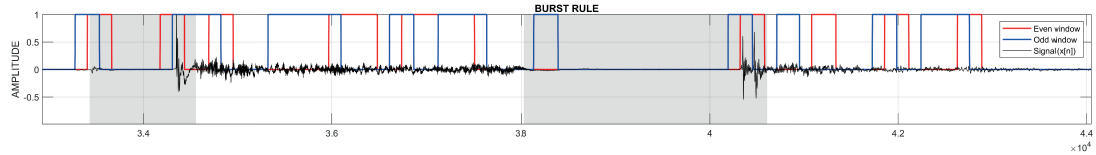

Figure 3.8: Burst Rule Detection

- **Combined Rule:** The presence of an unvoiced stop consonant is considered when all the three previously described characteristics coincide. A minimal period of silence occurs previous to the burst, the signal presents characteristics of an impulse-like signal by analysing the phase behaviour and a sudden increase of the signal energy is detected. Therefore, the combined rule signals an unvoiced stop consonant in the signal when all three characteristics are detected. Due to the fact that windows in the analysis overlap in time, two combination criteria are used:

  1. Criterion A, The logical combination of the three rules is achieved strictly for each window. Accordingly, the presence of an unvoiced stop consonant in the signal is verified when all three rules co-occur;
  2. Criterion B: The logical combination is performed for each voice segment. Therefore an unvoiced stop consonant is verified if all rules are verified in one of the two windows that the voice segment covers. This criterion allows the identification of an unvoiced stop consonant in the minimal duration of half a window.

  Figure 3.9 illustrates the combination of all three rules using criterion B.

Criterion B allows detecting unvoiced stop consonants when all conditions from the three rules are met for a signal region, even if that detection derives from adjacent windows. For example,

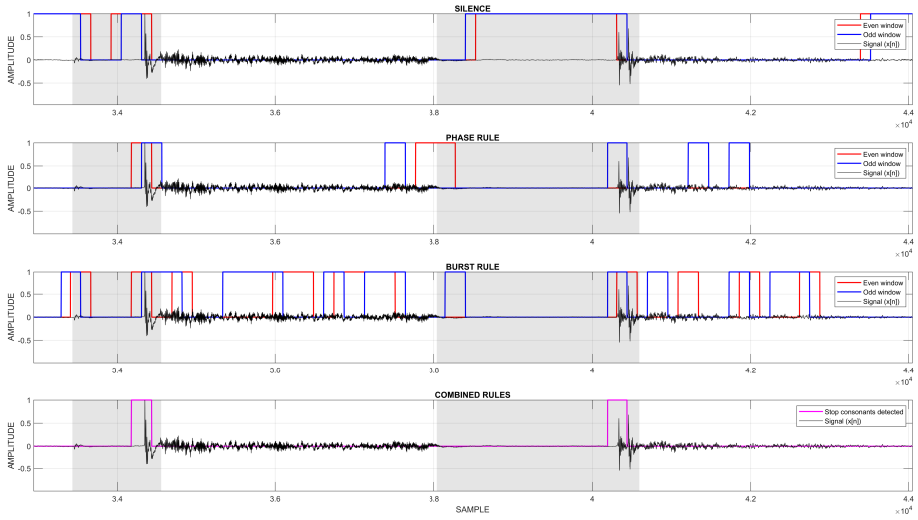

Figure 3.9: Combination of 3 rules using criterion B

consider the case where the silence rule is verified until window k, while the conditions relative to the burst and phase are only verified in window $k+1$ (the second half of window k corresponds to the first half of window $k+1$). In this setting, criterion A (single-window combination) would miss the presence of an unvoiced stop consonant due to all rules not being verified in the same window, as all three rules never coincide with this type of analysis. However, criterion B (voice segment combination) will detect the unvoiced stop consonant in the second half of window k (first half of window $k+1$) due to all rules being verified in at least one of the two windows of this voice segment set.

3.3 Performance

Testing the algorithm based on the DyNaVoiceR project database allowed the conclusion that a high percentage of unvoiced stop consonants is detected in isolated words. Criterion B managed to hit a superior performance, detecting 395 occurrences out of 420 existent, having a hit rate of 94%.

As expected, due to the rules implementation's nature, the algorithm does not consider the silence before the burst as being a part of the stop consonant, despite the manual segmentation considering it. This fact is not considered when calculating the ruleset's performance, due to the silence portion not needing voicing (done afterwards in the DyNaVoiceR project) for the stop to be perceptible in the end result. The silence rule allows preventing false positives, identified by the phase and burst rules, in the middle of words, namely vowels' diction, as shown in Figure 3.9.

Despite having a high performance, the algorithm is susceptible to identify wrong portions of the speech as unvoiced stop consonants, presenting many false positives, due to perturbations on the signal like external noise or by detecting words starting with consonants.

3.4 Summary

In this chapter, we addressed unvoiced stop consonants detection on the DyNaVoiceR database. The most prominent characteristics were studied and collected from it, allowing us to build an efficient and simple set of rules to detect them. It was described how these rules perform and the fundamentals behind each one. Therefore, the algorithm performance was examined using examples of its operation.

In the next chapter, a similar approach for detecting fricatives and sibilants is described, followed by the ruleset and performance obtained.

Chapter 4

Fricatives and Sibilants Detection

This chapter will focus on detecting fricatives (/f/, /v/) and sibilants (/s/, /ʃ/, /z/, /ʒ/), which are fricative consonants of higher amplitude and pitch. The distinction between fricatives and sibilants is made by realising that the sibilants have a broad peak of energy in higher frequencies, while fricatives tend to have a flatter spectrum similar to silence and a lower overall energy. Examples of both classes are shown during the next section.

Similarly to the previous chapter, a simple ruleset is created based on signal observations in fricative and sibilant DyNaVoiceR project database words.

4.1 Database

Once more, the database used is the one created specifically for the DyNaVoiceR project. As previously said, it includes recordings produced by 20 speakers (10 female and 10 male) in both voiced and whispered speech. The samples were recorded in the audio format .wav at a sampling rate of 22050 Hz and are supplemented with files including manual annotations, allowing to locate and identify each phoneme present in each word of the recordings.

An empirical examination of all unvoiced fricatives and sibilants in the DyNaVoiceR database was conducted to get some awareness of their signal characteristics. Table 4.1 shows all 18 words with voiced and unvoiced fricatives and sibilants present in the database, their phonetic sequence, and the stops' corresponding IPA symbol [30].

Furthermore, examples of signal and spectrogram of each phoneme studied in this chapter, in both voiced and whispered speech regarding speaker SPF01 of the DyNaVoiceR database, are shown in Figures 4.1 to 4.2 for fricatives and 4.3 to 4.6 for sibilants.

By observing the spectrograms, we can see an equilibrium of signal component in all frequencies for the class of phonemes we consider as fricatives for both whispered (right side) and voiced (left side) representations of the words. The same cannot be said for the sibilants, as there is a clear peak in higher frequencies, supporting our claims in treating both as different groups.

Words	Phonetic Sequence	IPA Symbol
<**ch**iba>	**C**V	ʃ
<la**j**e>	V**C**V	ʒ
<**f**a**c**e>	**C**V**C**V	f, s
<**v**a**z**e>	**C**V**C**V	v, z
<**s**ala>	**C**V	s
<**j**uba>	**C**V	ʒ
<**ch**ama>	**C**V	ʃ
<**v**ida>	**C**V	v
<a**ch**a>	V**C**V	ʃ
<**z**aro>	**C**V	z
<a**ss**a>	V**C**V	s
<**j**arra>	**C**V	ʒ
<**f**i**s**ga>	**C**V**C**C	f, ʒ
<a**s**a>	V**C**V	z
<ha**j**a>	V**C**V	ʒ
<a**ch**e>	V**C**V	ʃ
<**v**elho>	**C**V	v
<**v**iga>	**C**V	v

Table 4.1: Database of Fricatives and Sibilants

After analysing all the data gathered, a few preliminary conclusions were made:

- Fricatives are noise-like signals with a low broad energy distribution along all spectrum;
- Sibilants /s and /z have a broad peak of energy between 5000 to 9000 Hz;
- Sibilants /ʃ and /ʒ have a broad peak of energy between of 2000 to 5000 Hz.

These conclusions will be the foundation of the rules built in the next section.

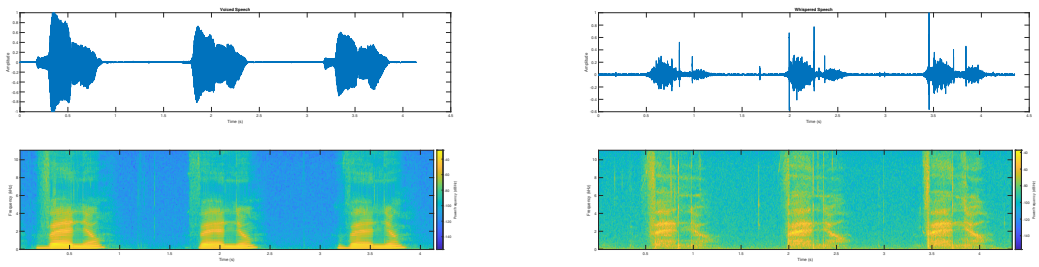

Figure 4.1: Recording of </v/elho> by SPF01

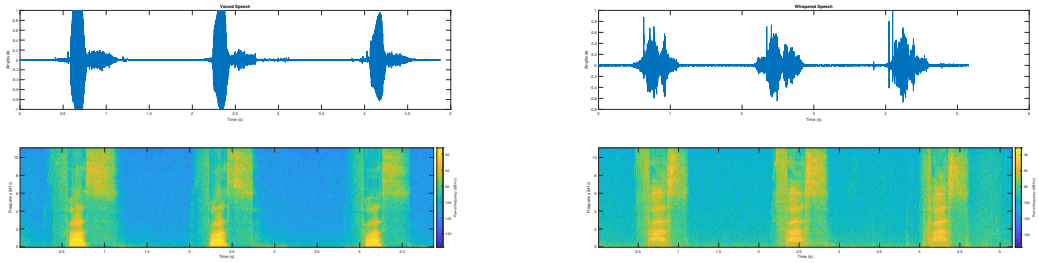

Figure 4.2: Recording of </f/ace> by SPF01

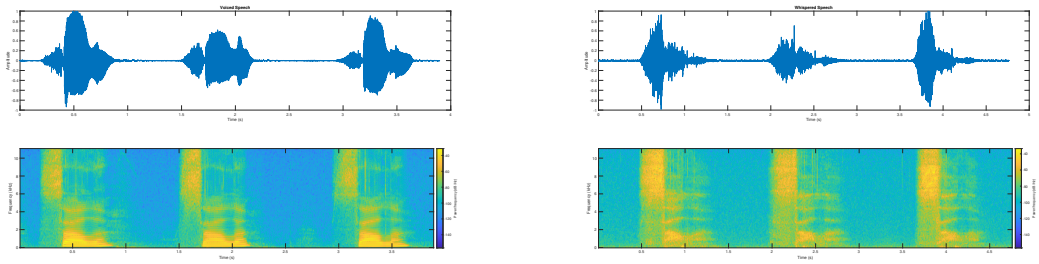

Figure 4.3: Recording of </s/ala> by SPF01

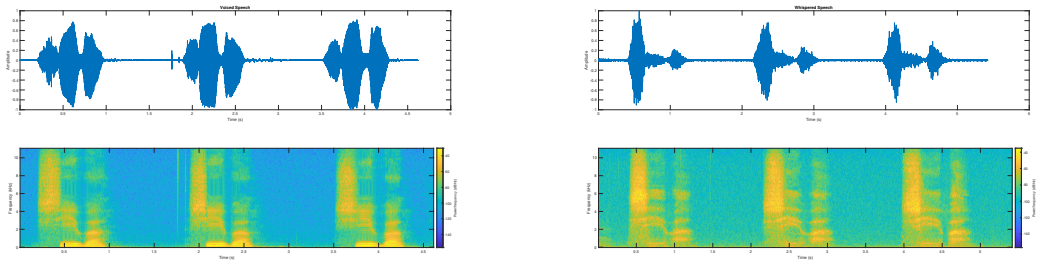

Figure 4.4: Recording of </ch/iba> by SPF01

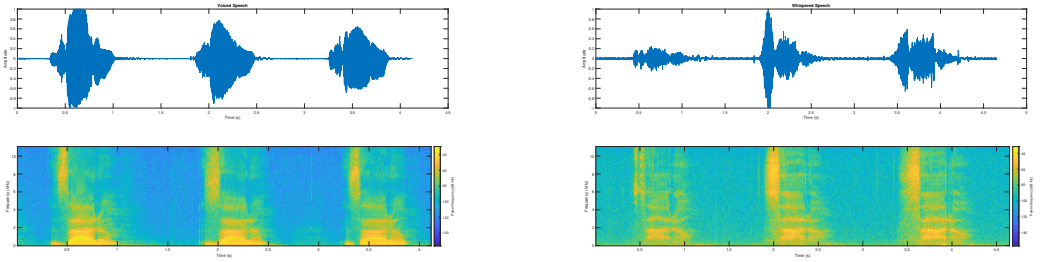

Figure 4.5: Recording of </z/aro> by SPF01

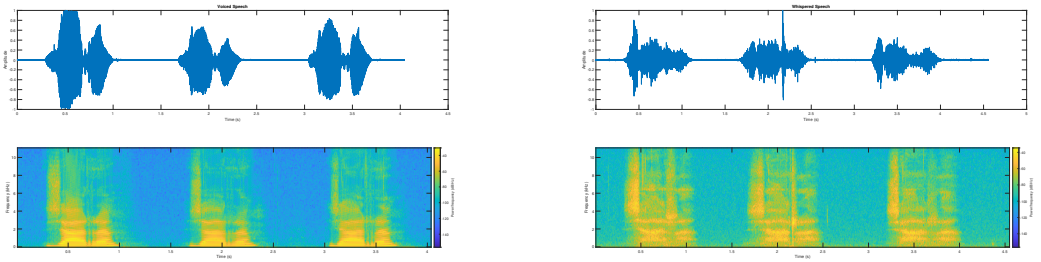

Figure 4.6: Recording of </j/arra> by SPF01

4.2 Methodology

This section presents the block diagram and the subset of rules used to detect both fricatives and sibilants. The rules will be supplemented with examples of their use and the reasoning behind them.

4.2.1 Block Diagrams

Figure 4.7 represents the fricatives detection algorithm's block diagram. The signal $x[n]$ represents the input signal, sampled at 22050 Hz of frequency, on which a sliding analysis window is applied with sinusoidal windows $w_s[n]$ with 256 samples and 50% signal overlap. We obtain the signal $x_W[n]$ for each window used to calculate the ODFT. Therefore, the signal obtained from the ODFT, $X_w[k]$, is used to calculate the Spectral Flatness Measure (SFM) and an energy ratio between the window energy above a defined frequency threshold and the total energy of the window. From there, both extracted features should be above defined thresholds for the signal segment to be considered a fricative.

Figure 4.8 represents the sibilants detection algorithm's block diagram. Once more, the process used to calculate the ODFT is replicated. From there, two energy ratios between the window energy above two defined frequencies and the window energy below the same frequencies thresholds are calculated. Once more, both extracted features should be above defined thresholds for the signal segment to be considered a sibilant.

It is noted that the rules are relatively similar. However, as explained before, the small differences between the two classes of fricatives justify the slight differences shown.

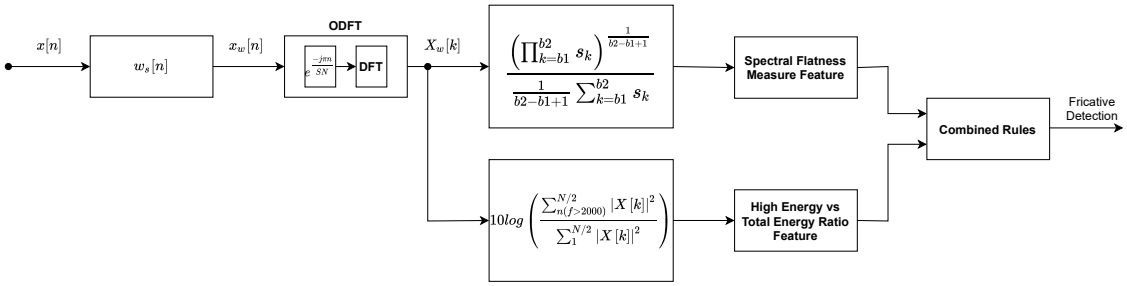

Figure 4.7: Fricatives detection algorithm's block diagram

4.2.2 Signal Characteristics and rules used

This subsection presents the subset of rules developed to detect two characteristics for fricatives and two characteristics for sibilants, respectively. The rules are based on the spectral flatness of the signal and variable energy ratios. Once the signal is processed, having a sliding window analysis with temporal overlap, the developed rules' output is identified in red, as Figure 4.9 shows, when the detection occurs. According to the manual annotation, the grey areas indicate a fricative or a sibilant presence, respectively.

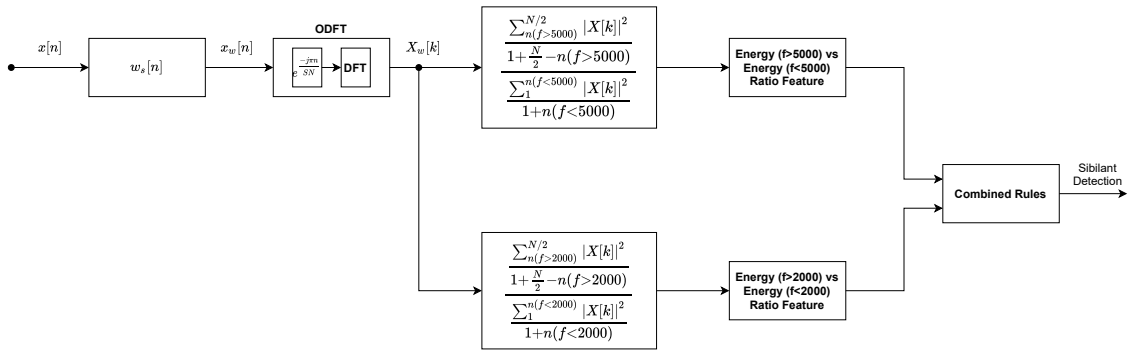

Figure 4.8: Sibilants detection algorithm's block diagram

4.2.2.1 Fricatives

- **Spectral Flatness Rule:** Fricatives are noise-like signals with a low broad energy distribution along all spectrum [32]. Due to this characteristic, the Spectral Flatness Measure (SFM) is suitable for this use case, as it calculates how flat the spectrum is. For instance, Equation (4.1) represents how the SFM is calculated and has a minimum threshold ($tr1$) that represents the flatness required for the signal be considered as a fricative;

$$\frac{\left(\prod_{k=b1}^{b2} s_k\right)^{\frac{1}{b2-b1+1}}}{\frac{1}{b2-b1+1}\sum_{k=b1}^{b2} s_k} > tr1 \tag{4.1}$$

  Where:

  - $b1$ & $b2$: band edges, in bins;
  - s_k : spectral value at bin k;
  - $tr1$: threshold 1;

- **High Frequencies Energy vs Total Energy Ratio Rule:** Despite the flatness of the signal indicating a fricative, noise signals and silence are also flat across the frequency spectrum, so to address the occurrence of false positives, a ratio between the spectrum energy for frequencies above 2000 Hz and the full spectrum energy is defined. This ratio helps eliminate false positives due to silence and noise containing most of their energy in lower energies, despite having flat spectrums. Therefore, Equation (4.2) shows how this ratio is calculated and how it needs to be above a minimum threshold ($tr2$) to reject the type of signals mentioned above.

$$10log\left(\frac{\sum_{n(f>2000)}^{N/2}|X[k]|^2}{\sum_{1}^{N/2}|X[k]|^2}\right) > tr2 \tag{4.2}$$

Where:

- N window size;
- X : audio signal in frequency domain;
- f : frequency
- $tr2$: threshold 2;

- **Combined Rule:** The presence of a fricative is considered when both of the above rules co-occur; the signal presents spectral flatness and has spectrum energy for frequencies above 2000 Hz over the total energy above ($tr1$ & $tr2$), respectively.

As shown in Figure 4.9, the energy rule's presence allows removing most of the false positives, even though some still happen due to noise above the noise floor being considered. Nevertheless, overall, the rules prove to be effective, as the results shown in the next section confirm just that.

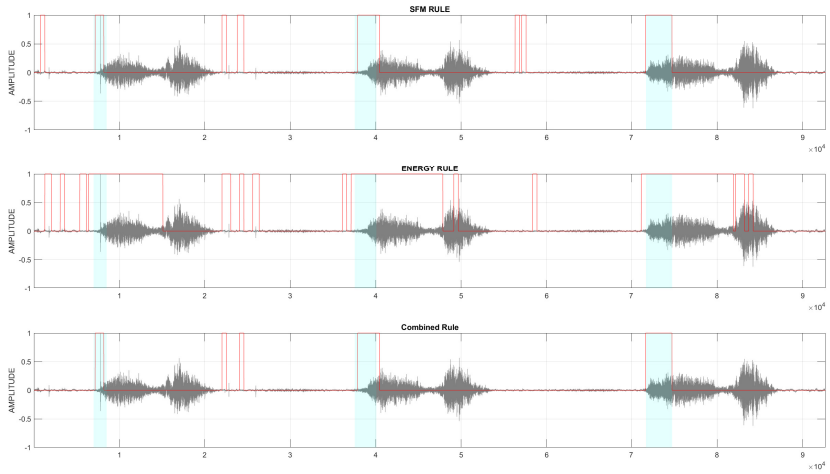

Figure 4.9: Fricative Rules for word </v/ida>

4.2.2.2 Sibilants

- **Frequencies Above 5000 Hz' Window Energy vs Frequencies below 5000 Hz' Window Energy Ratio Rule:** Sibilants /s/ and /z/, as shown by previous spectrograms and [33] have a defined broad peak of energy between 5000 to 9000 Hz. This rule applies this characteristic and compares the window's energy for frequencies above 5000 Hz with the window's energy for frequencies below that. Therefore, Equation (4.3) shows how this ratio is calculated and how it needs to reach a minimum threshold ($tr3$) to validate the rule.

$$\frac{\frac{\sum_{n(f>5000)}^{N/2}|X[k]|^2}{1+\frac{N}{2}-n(f>5000)}}{\frac{\sum_{1}^{n(f<5000)}|X[k]|^2}{1+n(f<5000)}} > tr3 \tag{4.3}$$

Where:

- N window size;
- X : audio signal in frequency domain;
- f : frequency
- $tr2$: threshold 2;

- **Frequencies Above 2000 Hz' Window Energy vs Frequencies Below 2000 Hz' Window Energy Ratio Rule:** Sibilants /ʃ/ and /ʒ/, as shown by previous spectrograms and [33] have a defined broad peak of energy between 2000 to 5000 Hz. This rule, once more, tries to apply this characteristic and compares the window's energy for frequencies above 2000 Hz with the window's energy for frequencies below that. Therefore, Equation (4.4) shows how this ratio is calculated and how it needs to reach a minimum threshold ($tr4$) to validate the rule.

$$\frac{\frac{\sum_{n(f>2000)}^{N/2}|X[k]|^2}{1+\frac{N}{2}-n(f>2000)}}{\frac{\sum_{1}^{n(f<2000)}|X[k]|^2}{1+n(f<2000)}} > tr4 \tag{4.4}$$

Where:

- N window size;
- X : audio signal in frequency domain;
- f : frequency
- $tr2$: threshold 2;

- **Combined Rule:** A sibilant presence is considered when one of the above rules occurs. Thus the signal has a ratio of spectrum energy for frequencies above 2000 Hz over spectrum energy for frequencies below 2000 Hz above a threshold ($tr3$). Or the signal has a ratio of spectrum energy for frequencies above 5000 Hz over spectrum energy for frequencies below 5000 Hz above a threshold ($tr4$).

As Figures 4.10 and 4.11 show, the rule used to detect sibilants /s/ and /z/ can detect all 4 sibilants studied, which is to be expected due to its range also containing the frequencies for /ʃ/ and /ʒ/. However, results shown in the next section prove there are some use cases where both combined can improve detection.

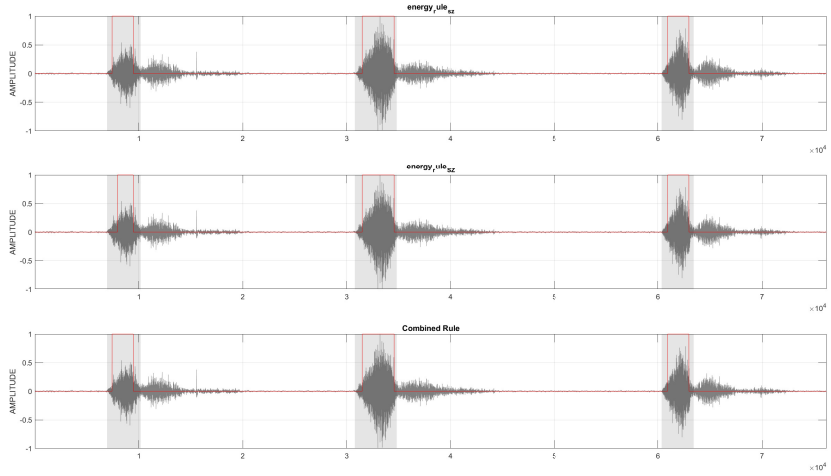

Figure 4.10: Sibilant Rules for word </s/ala>

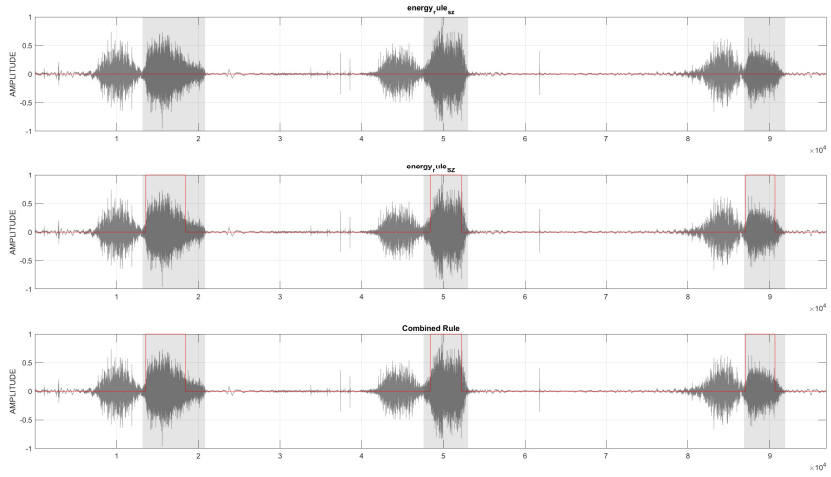

Figure 4.11: Sibilant Rules for word <la/j/e>

4.3 Performance

The algorithm testing over the DyNaVoiceR project database allowed concluding that 77% of fricatives are detected in 263 out of 341 occurrences. While the hit rate shows that the algorithm is not the most performant, it proves to be a good base for future improvements.

Additionally, the algorithm testing on the DyNaVoiceR project for the sibilants showed a hit rate of 81.6% in 266 out of 326 occurrences while using the rule based on the frequencies above and below 2000 Hz. Using both rules, we obtained a hit rate of 85.3% in 278 out of 326 occurrences. These results demonstrate that using both rules improves the hit rate, despite the second rule being redundant in most cases. This is the case due to, as explained before, the ratio of the first rule includes the bins of the frequencies of the second rule. The improvement happens in the rarer cases of sibilants having a higher percentage of their signal energy in frequencies above 5000 Hz, which the second rule is built to detect.

4.4 Summary

This chapter has presented the fricatives and sibilants on the DyNaVoiceR database. The most prominent characteristics were studied and collected from it, allowing us to build an efficient and simple set of rules to detect them. It was described how these rules perform and the fundamentals behind each one. From there, the algorithm performance for both phonemes was examined with examples of their operation.

The next chapter (5) presents the final work developed in this dissertation, the segmentation of all phonemes existent on the DyNaVoiceR database using HMMs.

Chapter 5

Hidden Markov Models

So far, we have focused on identifying a set of whispered speech's features for plosive and fricative phonemes. This chapter will focus on using those features as the base to build a stochastic approach to identify the studied phonemes and others of interest in whispered speech. Having that in mind, the Hidden Markov Models will be used as it is a widely used model in speech recognition. This chapter will explain its architecture, the type of HMM used and the results obtained using this approach.

5.1 HMM Parameters

The Hidden Markov Model (HMM) is a statistical Markov model in which the system being modelled is assumed to be a Markov process. A Markov process is a stochastic process that satisfies the Markov property [34] (usually characterised as memorylessness). Fundamentally, the current state occurrence does not depend on past events, even though the current hidden state prediction is based on past observations.

The characteristics of the model [35] are the following :

1. N, the number of states, $S = S_1, S_2, \cdots S_N$, in the model. Even though the states are hidden, there are many practical applications that there is a physical significance associated with each state. For example, in our study, each state will be a different phoneme of the Portuguese language. Usually, as is our case, each state can transition to any other existent in the model (e.g. an ergodic model), even though HMM allows any other states' interconnections.

2. M, the number of distinct observation symbols per state, $V = V_1, V_2, \cdots V_N$, which are the system's physical outputs. In our example, these will be the features extracted from the speaker signal.

3. The state transition probability distribution $A = A_{ij}$ where

$$a_{ij} = P\left[q_{t+1} = S_j \mid q_t = S_i\right], \quad 1 \leq i, j \leq N. \tag{5.1}$$

In our case, where any state can reach any other state in a single step, we have $a_{ij} \neq 0$ for all i, j. For any other type of HMMs, we would have $a_{ij} = 0$ for one or more (i, j) pairs.

4. The observation symbol probability distribution in state j, $B = B_j(k)$, where

$$b_j(k) = P[V_k \text{ at } t \mid q_t = S_j], \quad \begin{array}{l} 1 \leq j \leq N \\ 1 \leq k \leq M. \end{array} \tag{5.2}$$

5. The initial state distribution $\pi = \pi_i$ where

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N. \tag{5.3}$$

The above characteristics indicate that a complete HMM specification requires two model parameters (N and M), specification of observation symbols, and three probability measures A, B, and π. For convenience, we use the following compact notation

$$\lambda = (A, B, \pi) \tag{5.4}$$

to indicate the full parameter set of the model.

5.2 Methodology

In this section, the several steps needed to correct the HMM are explained and shown. They are comprised of the initial model parameters, the different features extraction and the model training.

5.2.1 Initial Model Parameters

As starting point, we divided the whispered words portion of the DyNaVoicerR database into a training and validation dataset to build the model. The validation dataset comprises 16 speakers (8 female and 8 male), corresponding to 80% of the total database words. In comparison, the validation portion comprises 4 speakers (2 female and 2 male), corresponding to the remaining 20% of the total database words. In order to choose which states the model is constructed by, we took the following considerations:

- The prominent phonemes in the database are stop consonants, fricatives and vowels.
- Limiting the number of states is beneficial to the model's hit rate when the database is relatively small, as it is the case.
- Grouping phonemes that rarely occur in the database into a single state helps increase that state's occurrence.

Having the previous assumptions in mind, we choose six states, as Table 5.1 shows.

Number	Name	Phonemes IPA Symbol
1	Silence	∅
2	Vowels	a,e,i,o,u,ɨ,ɐ,ɛ,ɔ
3	Stop Consonants	p,t,q,b,d,g
4	Fricatives	f,v
5	Sibilants	s,ʃ, z,ʒ
6	Others	m,n,ɲ,l,ʎ,ɾ,ʀ

Table 5.1: HMM states phonemes

We choose to separate the fricatives and sibilants phonemes for the reasons explained in the previous chapter. The state "others" includes all the consonants not contemplated in the previous states, with those being nasal, lateral and rhotic. For the observations, we based our feature set on the studies and tests conducted in the previous two chapters, as most of the words in the database contemplate those two phonemes categories. Together with the previous features, the Mel-frequency cepstral coefficients (MFCCs) [36] were added as a feature due to their recognised importance in speaker recognition [37] as well as phoneme recognition [38].

The observations features are the following:

- 13 MFCCs (Mel-frequency Cepstrum Coefficients);
- Current frame energy;
- Difference between current frame energy and the previous frame energy;
- Difference between current frame energy and the second previous frame energy;
- Spectral Flatness of the signal;
- Phase error

5.2.2 Feature Extraction

After determining the initial model parameters, determining the states transition matrix and observation symbol probability distribution are the next steps in implementing the HMM model. As such, the training dataset was sampled into windows of 1024 samples, with steps of 512 samples (50% overlap). Therefore, the signal is converted into the frequency domain by applying the ODFT (Odd Discrete Fourier Transform), a variant of the DFT (Discrete Fourier Transform), in which bin positions occur in the odd multiples of ($\frac{\pi}{N}$).

The next step is to calculate frame energy, SFM, phase error, and MFCC, which are calculated from the converted signal as described before. The various steps to get those features are presented in Figure 5.1.

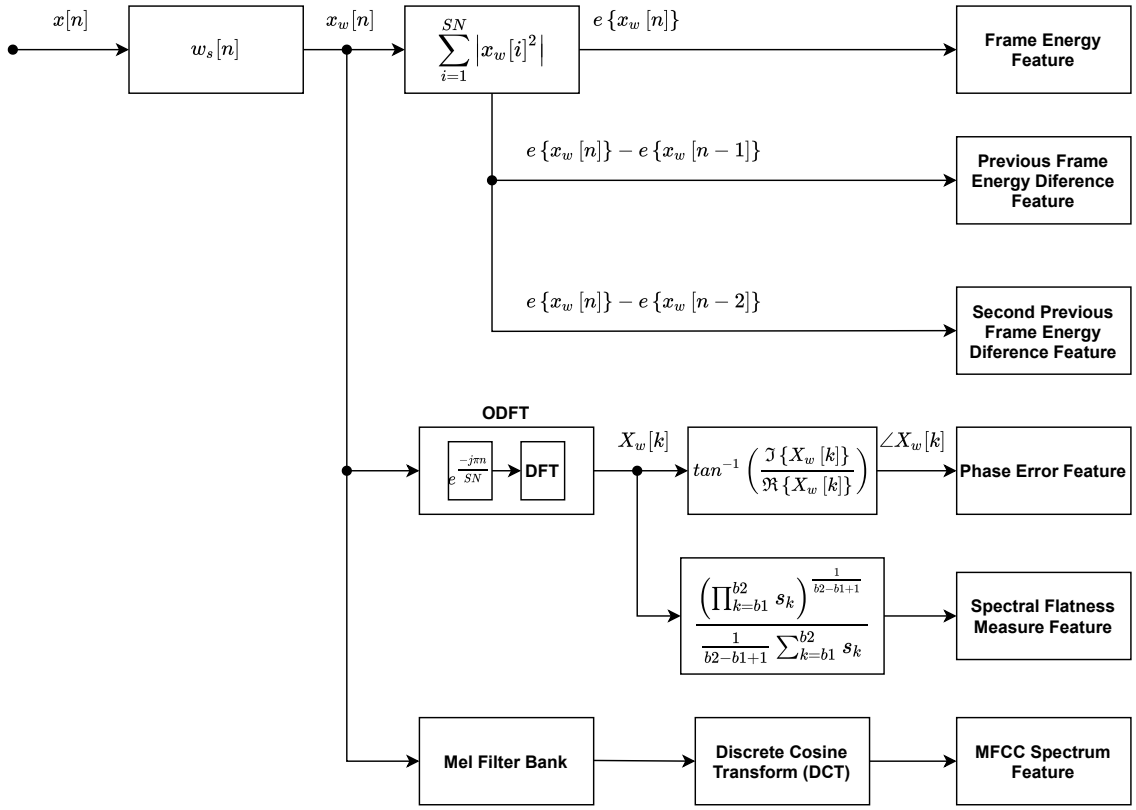

Figure 5.1: HMM' Feature Extraction Algorithm

Concurrently, each frame phoneme state is saved, as it will be the source of the states transition matrix by merely counting the number of times the state transitions to another or itself versus all frame occurrences.

5.2.3 Model Training

To estimate the model's parameters, at first, the HMM MATLAB toolbox was taken into account. The problem of the said toolbox was that it could only compute discrete observations, while in our case, we pretend to use decimal values obtained from the whispered speech signal (i.e. MFCC coefficients). After some research, we found the open-source toolbox PMTK3 [39], a framework for Matlab/Octave incorporating machine learning, graphical models, and Bayesian statistics, including the HMMs and the algorithms for the estimation of its parameters.

This toolbox uses the maximum a posteriori probability path estimator (MAP path estimator) to estimate the Markov chain's hidden path, whose properties derive from the Viterbi algorithm, an algorithm widely employed in coding theory, correction of intersymbol interference and text recognition [40].

The MAP path estimator [40] is applied to maximise the likelihood of the posterior probability distribution of features belonging to the correct state. That is achieved by employing an augmented optimisation objective that incorporates a prior distribution, usually called prior in machine learning,

that quantifies the additional information available through prior knowledge, which in this case is the states that the frames belong to while training the model.

In the end, the model returns variables compliant with HMM's compact Equation (5.4),which includes the state transition probability distribution Table 5.2, the initial state distribution Table 5.5 and the observation symbol probability distribution. The observation symbol probability distribution is modelled following a Gaussian distribution for each feature where each one has values of the mean and standard deviation (who describe their probability density function) for each state.

$$\pi = \{0.2857; 0.1429; 0.1429; 0.1429; 0.1429; 0.1429\} \tag{5.5}$$

From/To	Silence	Vowels	Stops	Fricatives	Sibilants	Others
Silence	0.9519	0.0104	0.0056	0.0096	0.0099	0.0125
Vowels	0.1226	0.7226	0.0614	0.0038	0.0464	0.0433
Stops	0.0008	0.3587	0.6335	0.0004	0.0004	0.0062
Fricatives	0.0011	0.3552	0.0011	0.6204	0.0011	0.0211
Sibilants	0.0481	0.1632	0.0149	0.0003	0.7667	0.0068
Others	0.0227	0.3303	0.0175	0.0048	0.0065	0.6182

Table 5.2: HMM's Transition Matrix

Following the HMM training, a process similar to that described in Figure 5.1 is conducted with the validation dataset. The feature extraction is carried so the model can "guess" the frame's hidden state through a dynamic programming algorithm. The algorithm used is the Viterbi algorithm, which finds the most likely sequence of hidden states by calculating the probability of all possible paths (trellis) that led to the current state and choosing the most probable.

5.3 Performance

This thesis's prototype simulations were performed using MATLAB R2020b 64- bit, through a processor AMD Ryzen 3700x with 8 cores and 16 threads. All audio files comply with the .wav format, using a sampling frequency of 22050 Hz. Without optimisations and in a development environment, the execution time by frame is inferior to 18 ms. This value is inferior to the frame rate, which at the sampling rate of 22050 Hz is 23 ms, confirming the viability of applying the algorithm in real-time.

Analysing the results of Table 5.3, which include all confusion matrix measures of all classes condensed into one table, we can see that the algorithm's overall performance is 80%. If the state "others", that involves phonemes with significant differences phonetically is ignored, a performance of 85% is achieved. Scrutinising each class independently, the stop consonants and fricatives classes' results show less hit rate than other classes, being around 50%.

In the case of the stop consonants, as explained before, to correctly perceive a stop consonant, we need to include a region of silence before the signal burst, which was taken into account in the annotation (ground truth). While that is correct perceptually, in terms of pure signal processing, we

Class	True Positives	False Positives	True Negatives	False Negatives
Silence	96%	18%	82%	4%
Vowels	81%	8%	92%	19%
Stop Consonants	52%	2%	98%	48%
Fricatives	45%	1%	99%	55%
Sibilants	69%	1%	99%	31%
Others	30%	2%	98%	70%
Total	80%	4%	96%	20%

Table 5.3: HMM Algorithm Classification Correctness

only need to do pure signal analysis where the voicing needs to happen, which means we only need to identify the burst.

As we can see in Figure 5.2, in the region ahead of the stop consonant burst, the HMM identifies the signal as being in silence, which is correct but will diminish the hit rate of the model (18% of false positives for the silence state) due to not being equal to the ground truth (which corresponds to the manual annotation).

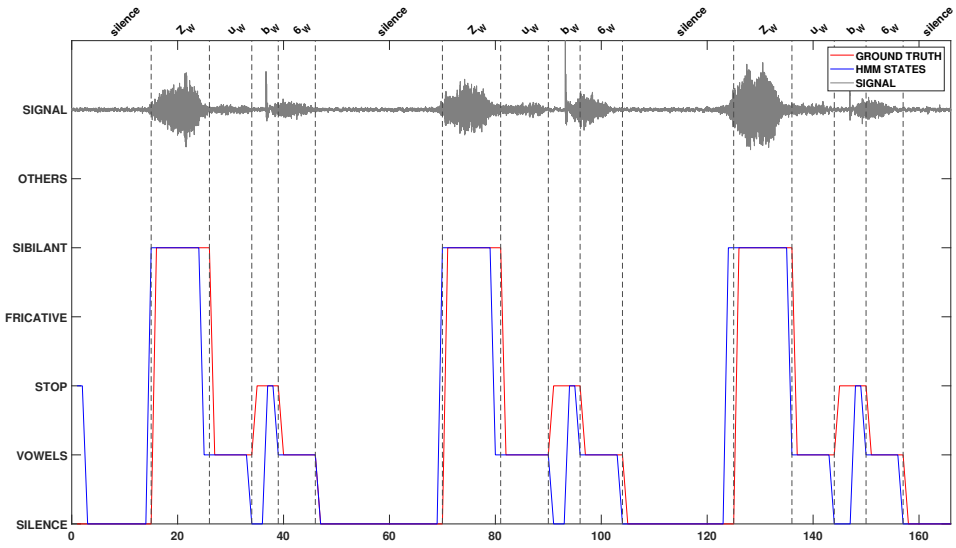

Figure 5.2: HMM performance for <juba>

In the fricatives' case, due to their nature of being noise-like signals without having defined peaks in their spectrogram, they are more sensitive to signal perturbations in the recordings. This can result in cases like the one in Figure 5.3, where a small signal perturbation resulted in a local peak, producing a misclassification of the frame as a stop consonant. The fricatives algorithm's performance could also be explained by not detecting them as fricatives in the phoneme's initial frames, wrongly misclassifying them as silence (18% of false positives for the silence state).

Furthermore, it can be noted that implementing an HMM utilising only MFCCs as features degrades performance by 5%, which complies with our thesis that using features learned from

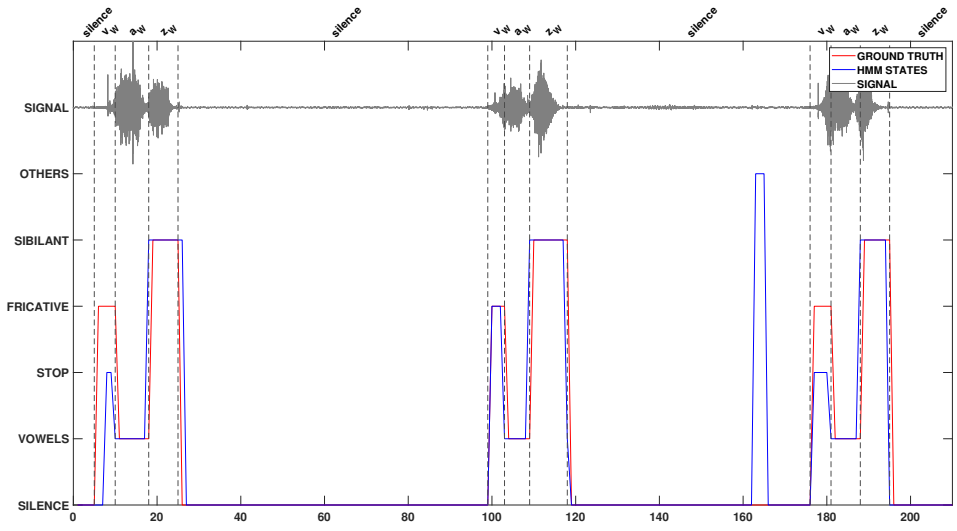

Figure 5.3: HMM performance for <vaze>

studying individual phonemes improves the performance of HMMs.

Despite these misclassifications, the overall performance of the HMM's can be considered satisfactory, indicated by the Figures 5.4 to 5.7 and the 85% of true positives if the state "others" is ignored. However, the performance could be improved further if the annotation took into account purely the signal characteristics and not the subjective nature of speech perception. Improvements to the variety and quantity of phonemes present in the DyNaVoiceR database would also prove valuable, as the current database is limited, especially when there is the need to build a statistical model that relies on the quantity of data fed to it as the performance of the state "others" (which includes several phonemes misrepresented in the database) show.

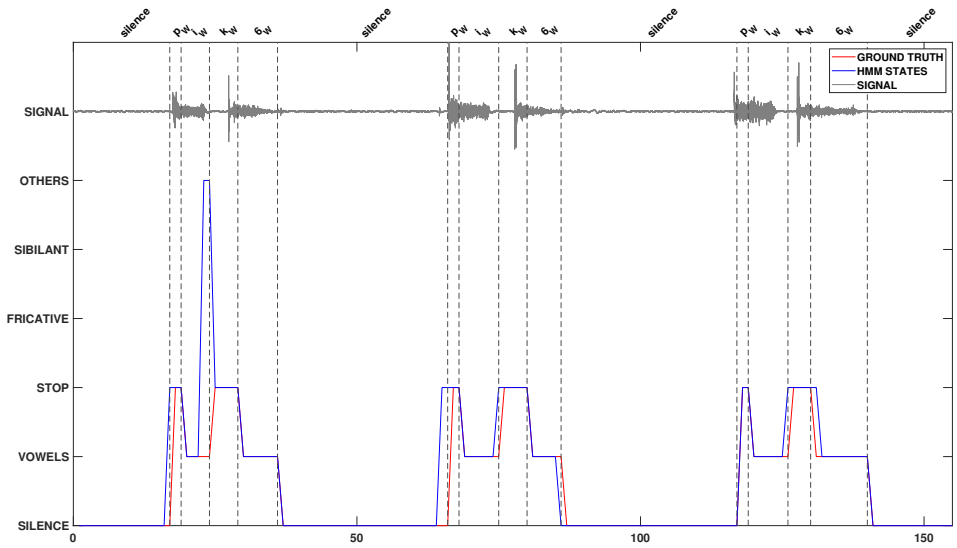

Figure 5.4: HMM performance for <pica>

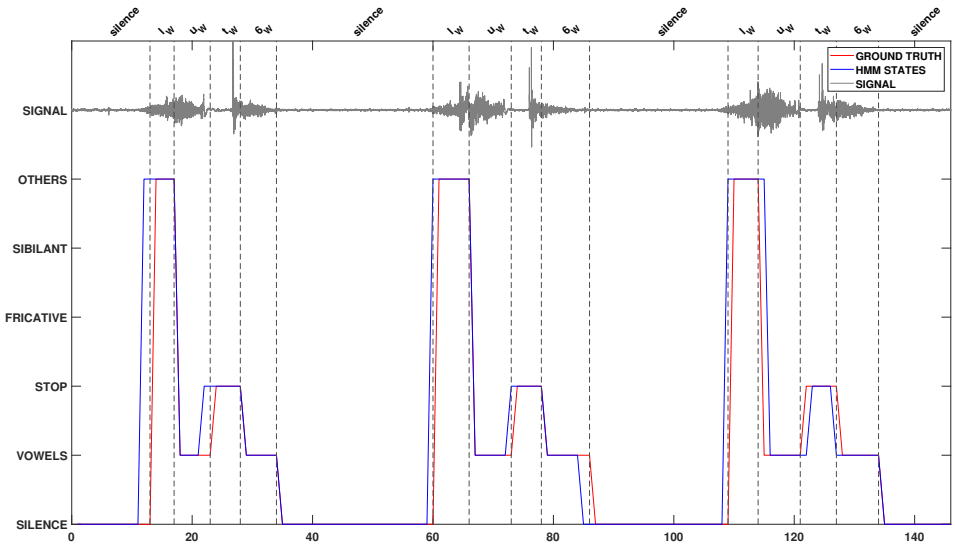

Figure 5.5: HMM performance for <luta>

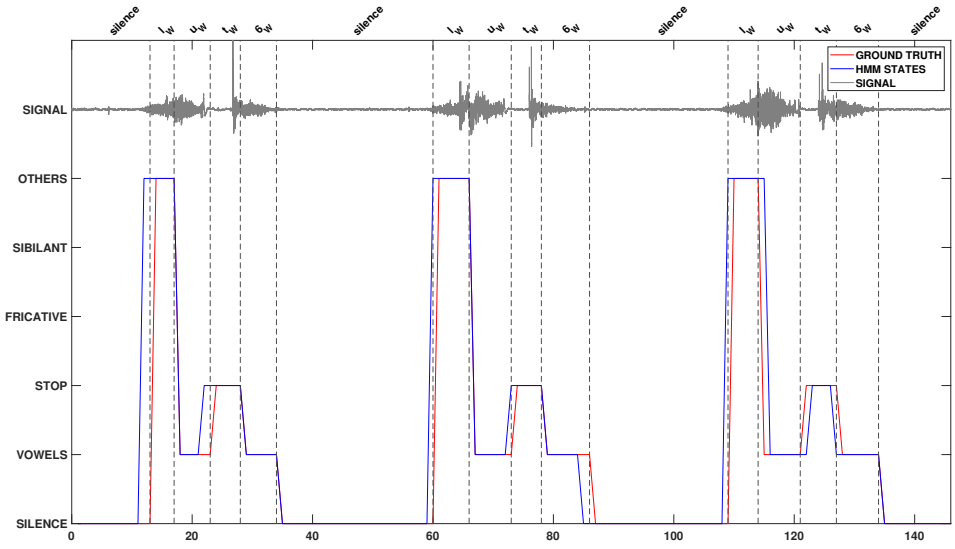

Figure 5.6: HMM performance for <fisga>

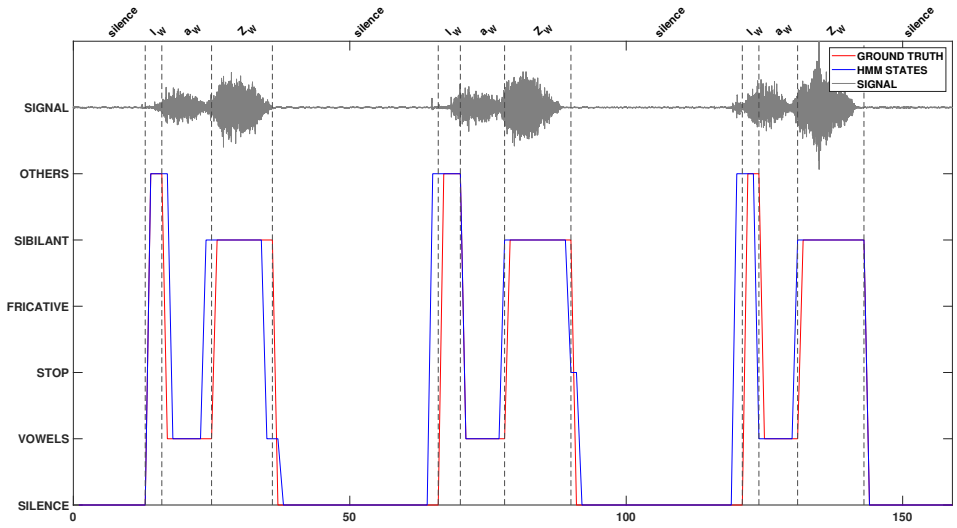

Figure 5.7: HMM performance for <laje>

5.4 Summary

This chapter first conducted a study of the HMMs operation and its different parameters. Afterwards, 6 states were chosen according to the DyNaVoiceR database's limitations, which were modelled using the Figure 5.1 features. Therefore, each feature was modelled for each state using a Gaussian Distribution by applying the MAP path estimator. Finally, using the Viterbi algorithm, the model was trained and obtained the Table 5.3 results. We managed to hit an 85% hit rate for the classes chosen (discarding state "others") in whispered speech while having some limitations in the database. Further optimisations in the annotation of phonemes and quantity and variety of phonemes in the database could lead to an enhanced hit rate, proving the viability of the approach used.

The following and final chapter (6) displays the conclusions obtained in this dissertation and the future works that may derive from it.

Chapter 6

Conclusions and Future Work

Many individuals are constrained to only using the whispered speech due to some aphonia or dysphonia, impacting their lives personally and professionally. These include Paradoxical Vocal Fold Movement, Vocal Fold Nodules and Polyps, Spasmodic Dysphonia and Vocal Fold Paralysis. The solutions currently in use for whispered speech to voiced speech conversion do not include a practical approach to implement them in the patient, nor do they have the practicality of producing results in real-time. As a result, the DyNaVoiceR project in which this thesis is included arose to solve these limitations. Moreover, this is a pioneer project in Portugal, where no solution other than the old electrolarynx is available for dysphonic patients.

In this project's scope, this thesis emerged to focus on implementing whispered speech segmentation for the Portuguese language, which is proven to facilitate voicing whispered speech.

Therefore, the following was achieved in this dissertation:

- **Segmentation of unvoiced stop consonants in whisper using simple rules:** Analysing the stop consonants signal characteristics, three rules that use the window energy, comparisons between consecutive windows' energies and the phase error were built. A hit rate of 94% was obtained;
- **Segmentation of fricatives in whisper using simple rules:** Analysing the fricatives signal characteristics, two rules that use the signal's spectral flatness and a ratio between different frequency bands were built. A hit rate of 77% was obtained;
- **Segmentation of fricatives in whisper using simple rules:** Analysing the sibilants signal characteristics, two rules that use two different ratios between different frequency bands were built. A hit rate of 85% was obtained;
- **Implementation of an automatic phoneme recogniser for silence, vowels, stop consonants, fricatives and sibilants utilising Hidden Markov Models:** Dividing the DyNaVoiceR's database words phonemes into 6 states and extracting 18 features (13 MFCCs, 3 frame energy features, signal's spectral flatness, and phase error), the HMM was trained using the MAP path estimator and the hidden states were extracted using the Viterbi Algorithm. A performance of 85% was obtained.

The results obtained are auspicious and allow execution in real-time, proven by the HMM model's execution time of 18 ms, which is inferior to the requirement of 23 ms at the sample rate of 22050 Hz. Furthermore, several doors were opened to drive the project further, on which the future work of this dissertation will focus.

For instance, the following is proposed:

- **Implementation of a noise-filter for the input signal:** As shown by the results of the simple rules algorithms and HMMs, the occurrence of noise degrades the efficiency of the algorithm by generating false positives where only silence exists due to whispered speech being sensitive to perturbations, as it is naturally a quieter manner of speaking.
- **Analyse different phonemes individually:** As shown by the study of unvoiced stop consonants, fricatives and sibilants, the simple ruleset created with the detailed study of each phoneme individually allowed creating the knowledge necessary to build the features used in the HMMs. Therefore, studying further phonemes would allow discovering additional features that could be used to improve the HMMs' model.
- **Increase the quantity and variety of phonemes in the database:** HMMs results showed that states with a limited number of samples for specific phonemes deteriorate the model's ability to learn the features associated with them. A wider variety of phonemes would also allow using more phoneme states, improving the model's versatility when used for segmentation purposes.
- **Implement an adaptive Noise-floor:** As the results were obtained using a database where the recording environment was controlled, the algorithms and the HMMs lack the ability to adapt to different environments where the noise-floor may change with time. Therefore, by adapting the energy level in which silence is detected to the average of the noise-floor in a specific pre-defined time interval, the algorithm's would be more versatile.
- **Thresholds that use the frame energy should be compared to the maximum frame energy:** Once more, as the database's recording environment was controlled, the algorithms could be overfitted to the recordings' loudness. By comparing the current frame's energy with its maximum, the algorithm's versatility could improve;
- **Utilise an external database to extract phonemes features prior distributions for HMM training:** If an external database is utilised to extract prior distributions for the phonemes studied, we can get a starting point closer, while training, to the best of our database, avoiding getting stuck in local minimums.

References

[1] Gustavo Andrade-Miranda. *Analyzing of the vocal fold dynamics using laryngeal videos*. PhD thesis, 06 2017. doi:10.20868/UPM.thesis.47122.

[2] Georgios Kouroupetroglou and Georgios Chrysochoidis. Formant tuning in byzantine chanting. *Int. Conference Sound and Music Computer*, 07 2014.

[3] L. Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[4] Jamie A. Koufman and Christie Block. Differential diagnosis of paradoxical vocal fold movement. *American Journal of Speech-Language Pathology*, 17(4):327–334, 2008. doi:10.1044/1058-0360(2008/07-0014).

[5] CD García Alvarez, ME Campos Bañales, D López Campos, J Rivero, B Pérez Piñero, and D López Aguado. Polyps, nodules, and reinke edema. an epidemiological and histopathological study. *Acta otorrinolaringologica espanola*, 50(6):443—447, 1999. URL: http://europepmc.org/abstract/MED/10502695.

[6] M J Aminoff, H H Dedo, and K Izdebski. Clinical aspects of spasmodic dysphonia. *Journal of Neurology, Neurosurgery & Psychiatry*, 41(4):361–365, 1978. URL: https://jnnp.bmj.com/content/41/4/361, arXiv:https://jnnp.bmj.com/content/41/4/361.full.pdf, doi:10.1136/jnnp.41.4.361.

[7] Hamid Daya, Asaad Hosni, Ignacio Bejar-Solar, John N. G. Evans, and C. Martin Bailey. Pediatric Vocal Fold Paralysis: A Long-term Retrospective Study. *Archives of Otolaryngology–Head Neck Surgery*, 126(1):21–25, 01 2000. URL: https://doi.org/10.1001/archotol.126.1.21, arXiv:https://jamanetwork.com/journals/jamaotolaryngology/articlepdf/404068/ooa90018.pdf, doi:10.1001/archotol.126.1.21.

[8] Hanjun Liu and Manwa L. Ng. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx*, 34(3):327 – 332, 2007. URL: http://www.sciencedirect.com/science/article/pii/S0385814606001751, doi:https://doi.org/10.1016/j.anl.2006.11.010.

[9] J. Wang, A. Samal, J. R. Green, and F. Rudzicz. Sentence recognition from articulatory movements for silent speech interfaces. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4985–4988, March 2012.

[10] Kenneth N. Stevens. *Acoustic phonetics*. NetLibrary, Inc., 1999.

[11] J. Coleman, E. Grabe, and Bettina Braun. Larynx movements and intonation in whispered speech. 01 2002.

[12] Ian Vince McLoughlin. *Audio analysis*, page 195–222. Cambridge University Press, 2016. doi:10.1017/CBO9781316084205.009.

[13] Xuedong Huang, Alejandro Acero, and Hsiao-Wuen Hon. *Spoken language processing: a guide to theory, algorithm, and system development*. Prentice Education Taiwan, 2005.

[14] Bernard Gold, Nelson Morgan, Dan Ellis, and Bourlard Herve. *Speech and audio signal processing: processing and perception of speech and music*. Wiley, 2011.

[15] Gunnar Fant. *Analys av de svenska konsonantljuden: talets allmanna svangningsstruktur*. L.M. Ericsson, 1949.

[16] Gunnar Fant. Voice source modeling. *The Journal of the Acoustical Society of America*, 73(S1), 1983. doi:10.1121/1.2020533.

[17] Tom Bäckström and Okko Räsänen. Introduction to speech processing. *Aalto University*, 2019.

[18] Maria Helena Mira. Mateus, Fale Isabel, and Freitas Maria Joao. *Fonetica e fonologia do portugues*. Universidade Aberta, 2005.

[19] Stephen Cassidy and Jonathan Harrington. The place of articulation distinction in voiced oral stops: Evidence from burst spectra and formant transitions. *Phonetica*, 52(4):263–284, 1995. doi:10.1159/000262182.

[20] E. Eide, J.r. Rohlicek, H. Gish, and S. Mitter. A linguistic feature representation of the speech waveform. *IEEE International Conference on Acoustics Speech and Signal Processing*, 1993. doi:10.1109/icassp.1993.319347.

[21] Dennis H. Klatt. Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, 18(4):686–706, 1975. doi:10.1044/jshr.1804.686.

[22] Sharlene A. Liu. Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America*, 96(5):3227–3227, 1994. doi:10.1121/1.411152.

[23] Hari Krishna Vydana and Anil Kumar Vuppala. Detection of fricatives using S-transform. *The Journal of the Acoustical Society of America*, 140(5):3896–3907, 2016. doi:10.1121/1.4967517.

[24] D. Ruinskiy, N. Dadush, and Y. Lavner. Spectral and textural feature-based system for automatic detection of fricatives and affricates. In *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pages 771–775, 2010. doi:10.1109/EEEI.2010.5662106.

[25] Anita Wagner, Mirjam Ernestus, and Anne Cutler. Formant transitions in fricative identification: The role of native fricative inventory. *The Journal of the Acoustical Society of America*, 120(4):2267–2277, 2006. doi:10.1121/1.2335422.

[26] K. Geetha and E. Chandra. Automatic phoneme segmentation of tamil utterances. In *2015 International Conference on Advanced Computing and Communication Systems*, pages 1–4, 2015. doi:10.1109/ICACCS.2015.7324062.

[27] M. A. Khawaja and N. G. Haider. Segmentation of sindhi speech using formants. In *2007 IEEE International Conference on Signal Processing and Communications*, pages 796–799, 2007. doi:10.1109/ICSPC.2007.4728439.

[28] M.J.F. Gales and Steve Young. The application of Hidden Markov Models in speech recognition. *Foundations and Trends in Signal Processing*, 1:195–304, 01 2007. doi:10.1561/2000000004.

[29] S. J. Leow, E. S. Chng, and C. Lee. Language-resource independent speech segmentation using cues from a spectrogram image. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5813–5817, 2015. doi:10.1109/ICASSP.2015.7179086.

[30] Caroline Smith. Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet (1999). *Phonology*, 17:291–295, 08 2000. doi:10.1017/S0952675700003894.

[31] Aníbal ferreira and Deepen Sinha. Frequency-domain parametric coding of wideband speech–a first validation model. *Journal of the Audio Engineering Society*, October 2015.

[32] Anibal Ferreira. Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information. *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, 2016. doi:10.1109/isivc.2016.7893980.

[33] Teixeira de Jesus Luis Miguel. *Acoustic phonetics of European Portuguese fricative consonants*. PhD thesis, 2001.

[34] Paul A. Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley Sons, Inc., 2017.

[35] L.r. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi:10.1109/5.18626.

[36] Tom Bäckström. Cepstrum and MFCC. URL: https://wiki.aalto.fi/display/ITSP/CepstrumandMFCC.

[37] Roma Bharti and Priyanka Bansal. Real time speaker recognition system using MFCC and vector quantization technique. *International Journal of Computer Applications*, 117(1):25–31, 2015. doi:10.5120/20520-2361.

[38] M. D. Skowronski and J. G. Harris. Increased filter bandwidth for noise-robust phoneme recognition. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–801–I–804, 2002. doi:10.1109/ICASSP.2002.5743839.

[39] Matt Dunham and Kevin Murphy. Pmtk3, Dec 2011. URL: https://github.com/probml/pmtk3.

[40] A. Caliebe. Properties of the maximum a posteriori path estimator in Hidden Markov Models. *IEEE Transactions on Information Theory*, 52(1):41–51, Jan 2006. doi:10.1109/TIT.2005.860425.