Parkinson's Disease Progression Assessment From Speech

Tomás Arias Vergara

Supervisor: Prof. Dr-Ing. Juan Rafael Orozco-Arroyave

Faculty of Engineering
Universidad de Antioquia

This dissertation is submitted for the degree of
Master of science

Grupo de Investigación en
Telecomunicaciones Aplicadas-GITA

July 2017

Acknowledgements

I wish to express my deep gratitude to all the people that supported me from the start of this work and made possible for me to accomplish several goals.
First of all I want to thank my advisor Prof. Dr-Ing Juan Rafael Orozco, which gave me the opportunity to explore the world of academic research. His guidance, encouragement, and invaluable support have been very important to me, not only for the development of this work, but also to identify several of my career goals. I can only offer my sincere appreciation for the learning opportunities provided by Rafa.
I also want to thank my colleagues from the GITA research group. Camilo for all the help I received from him during the development of this work. He has been the person with whom I can develop new ideas, discuss the results of many experiments, and share a beer once in a while. Thanks also to Elkyn, Tatiana, and Nicanor for their academic support in the first stage of this work. I would want to thank also to Prof. Dr-Ing Elmar Nöth for the opportunity to learn as part of the Pattern Recognition Lab. of the Friedrich Alexander Universität Erlangen-Nürnberg. From the first moment I arrived in Germany he was very supportive in both academic and personal ways and for that I will be always grateful.
To my family for their support during all my academic life. Especially to my mom for her wisdom and guidance. She have always gave me reasons to keep going, go beyond my limits, and dream for the best. To the Colombian Administrative Department of Science, Technology and Innovation (COLCIENCIAS) for supporting this study through the young researches and innovators program in 2015 and the COLCIENCIAS project # 111556933858 titled “Analysis of the discriminant capacity of phonation, articulation and prosody features from patients with Parkinson’s disease on preclinic and advanced stages for the development of computer aided tools for supporting the diagnosis and monitoring of the patients”.
And last but not least, I want to thank the patients and volunteers of the Fundalianza Parkinson Colombia. Without their help and willingness to collaborate in this work none of this could have been possible. Thanks to them for letting me be part of their group.

Abstract

In recent years the interest in the analysis of speech of Parkinson's disease patients have been increasing. Most of the studies are focused on developing computer aided tools for the unobtrusive monitoring of the disease. Different approaches have been proposed in order to detect several voice problems in Parkinson's patients. However, a relative low number of the contributions are focused on tracking the state of the disease over the time. In this thesis is proposed a new methodology to assess the progression of the disease per speaker. The speech of the patients is modeled individually considering speech recordings captured in different sessions. Voice impairments are analyzed considering a particular model that reflects articulatory problems of the patients. Different measures are considered to detect changes in the speech of the patients and quantify the state of the disease. The estimated measures are compare to the neurological state of the patients assessed by expert clinicians according to a standard rating scale. Additionally, the features used to model the speech deficits of the patients are validated by predicting their neurological state according to the Movement Disorder Society-Unified Parkinson's Disease Rating Scale. The proposed approach shows that is possible to track the disease progression from speech with a relative high correlation value.

Table of contents

Chapter 1

Introduction

1.1 Motivation

Parkinson disease (PD) is the second most prevalent neurodegenerative disease after Alzheimer's [1]. About 2 % of the people older than 65 years have PD [2]. According to a study on 10 of the world's most populous nations, in 2005 the number of people with PD was between 4.1 million and 4.6 million and it is estimated that this number will rise to 8.7 - 9.3 millions in 2030 [3]. For the case of Colombia, the prevalence of the disease is about 172 cases per each 100.000 inhabitants [4]. The region with the highest prevalence is Antioquia with 30 cases per each 100.000 inhabitants [5].

Parkinson's disease is a neurodegenerative disorder characterized by the progressive loss of dopaminergic neurons in the substantia nigra of the midbrain [6]. The primary motor symptoms of PD include tremor, slowness, rigidity of the limbs and trunk, and postural instability. PD also affects the muscles and limbs involved in the speech production process [7]. Some of the voice impairments include fast speech, soft voice, hoarse quality of voice, and others. About 90 % of PD patients experience changes in voice that affects their communication ability [2]. Many of the symptoms are controlled with medication, however there is no clear evidence indicating positive effects of those treatments on the speech impairments [8]. There is also evidence that shows that a proper speech therapy combined with the pharmacological treatment can improve the communication ability of PD patients [9].

The progression of PD and the symptoms experienced by some patients vary from one person to another. The neurologist relies on medical history, physical and neurological examinations to assess the patients. However, the motor skills of the patients with PD are impaired, thus to visit a hospital to perform medical screenings and/or assessments is not a straightforward task for them [10]. Additionally, the diagnosis and monitoring of PD

symptoms is time-consuming and expensive. For these reasons there is an increasing interest from the research community to develop computer aided systems for monitoring of PD patients. The continuous monitoring of PD patients could help to make timely decisions regarding their medication and their therapy.

1.2 State-of-the-art

The disease severity is evaluated by neurologist experts by means of several tests and scales. One of them is the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [11]. This is a perceptual scale used to assess motor and non-motor abilities of PD patients. The total MDS-UPDRS scale is divided into four sections. Most of the works concerning speech assessment of patients considers only the third section (MDS-UPDRS-III) which consists of the evaluation of the motor capabilities.
In this section different studies focused on the detection and tracking of PD from speech are reviewed.

1.2.1 Monitoring of PD from speech

In [12] the authors presented a methodology to predict the disease progression from speech signals. Recordings of sustained vowel phonations were modeled using several acoustic measures including jitter, shimmer, Noise to Harmonic Ratio (NHR), Harmonic to Noise Ratio (HNR), Relative Amplitude Perturbation (RAP), Period Perturbation Quotient (PPQ), Amplitude Perturbation Quotient (APQ), Recurrence Period Density Entropy (RPDE), Detrented Fluctuation Analysis (DFA), and Pitch Period Entropy (PPE). The UPDRS scores were predicted using three linear regression techniques: Least Squares (LS), Iteratively Re-weighted Least Squares (IRLS), and Least Absolute Shrinkage and Selection Operator (LASSO). Additionally, the authors predicted the UPDRS scores using the Classification And Regression Trees (CARTs) approach. The speech of 42 PD patients (28 male, 14 female) was recorded once per week during six months. Neurologist experts evaluated the patients three times during the study. To obtain the weekly UPDRS scores, the authors used a piecewise linear interpolation. The performance of the regression techniques is evaluated using the Mean Absolute Error (MAE). The authors reported that the best results were obtained with the CARTs approach with a MAE of 7.5 points in the predictions of the total value of the UPDRS scale. Further, the UPDRS-III was predicted with a MAE of 6 points. The novelty of the method is the disease severity assessment from speech. However, it seems like the authors were not aware of the speaker independence in their experiments because they mixed

recordings of the test patients with the recordings of the train set. Thus, the reported results are highly optimistic and biased.

The progression of speech impairments in a longitudinal study is presented in [13]. The speech of 80 PD patients (48 male, 32 female) was recorded from 2002 to 2012 in two different recording sessions. The time between the first and second session ranged from 12 to 88 months. A control group of 60 healthy persons (30 male, 30 female) was also considered. The participants were asked to read a text and to produce a sustained phonation of the vowel /a/. In both sessions the patients were assessed by neurologist experts according to the UPDRS-III. The audio signals were perceptually evaluated by two of the authors (S. Skodda and W. Grönheit). Four aspects of speech were considered: voice, articulation, prosody, and fluency. These aspects are used by the authors to describe dysarthria, i.e., motor speech disorders. Additionally, an acoustic analysis is performed to describe these speech aspects. Voice was modeled with a set of features that includes jitter, shimmer, NHR, and average pitch. For articulation the Vowel Articulation Index (VAI) and the percentage of pauses within polysyllabic words are considered. Prosody is analyzed with the estimation of the standard deviation of the pitch. Fluency was evaluated considering the Net Speech Rate (NSR) and the pause ratio. To assess the progression of speech and voice impairments the authors performed three experiments: In the first experiment, they calculated the Pearson's correlation of the scores of the perceptual evaluation with the speech item of the UPDRS scale and the acoustic features. In the second experiment they compared the PD group and the healthy speakers group (Shapiro-Wilk test) in the first session. In the third experiment they compared the extracted features from the PD group estimated in the first session and the second session. The authors reported significant correlations between the perceptual scores and the UPDRS item. Additionally, the authors report that there were significant differences between the PD patients and the HC group in both, the perceptual scores and the acoustic features. For the third experiment the authors found significant differences for shimmer, NHR, NSR, pause ratio, and VAI when features extracted from the first session of the PD patients are compared with the same features calculated upon the second session. According to the authors, the results are not conclusive due to some methodological limitations like the different intervals in the recording sessions.

A study for the monitoring of PD progression is also presented in [14]. The authors recorded a total of four male patients every week during one month in four recording sessions. Speech recordings of 100 healthy speakers (50 male, 50 female) were also considered in their study. The authors estimated several features to describe speech disorders as consequence of neuromotor instability in phonation, which is defined as the process of producing vocal sounds from the vibration of the vocal folds. Sustained phonations of the vowel /a/ was modeled

using different features to describe tremor (first order pivoting coefficient, physiological tremor amplitude, neurological tremor amplitude, flutter amplitude, and global tremor), perturbation of the vocal folds (pitch, jitter, shimmer, and NHR), and biomechanical phonation impairment (vocal folds body mass, body stiffness, cover mass, cover stiffness, adduction defect, and glottal gap). The authors used two methods to estimate features of tremor and biomechanical impairments: (1) a vocal tract inversion by a lattice adaptive filter and (2) biomechanical inversion of a 2-mass model of the vocal folds. Features from the 50 male healthy controls (HC) were used as the baseline to describe the normal state of the speech. During the month of the recordings, the patients continued their pharmacological treatment and received speech therapy. Each patient was evaluated according to the Hoehn and Yhar (H&Y) scale. The suitability of the features used to describe phonation impairments was evaluated by means of a metric defined as the weighted sum of the extracted features as a function of a sigmoid that ranges from 0 to 5. The aim of this metric is to estimate the relationship between the H&Y scores on each recording session and the phonation features estimated for vocal fold perturbation, tremor, and biomechanical impairment. According to the authors the most relevant features are jitter, vocal fold body mass, body stiffness, adduction defect, physiological and neurological tremor amplitude, flutter amplitude, and global tremor. Additionally, the authors report that tremor and biomechanical features evolve differently with the treatment. However, the authors stated that it is necessary to define different time intervals between evaluations to obtain more conclusive results. Similarly, in [15], the authors proposed the Log Likelihood Improvement Ratio (LLIR) as a metric to compare speech recording of eight male PD patients captured in four recording sessions. The patients followed a pharmacological treatment and received speech therapy. The aim of the study was to detect changes in the voice before and after treatment using the same features described before. The authors report that the LLIR is a good metric to detect changes in phonation when the patient is under treatment (or therapy). Although the authors detected changes in phonation measures it is not clear whether the same approach is able to detect changes in the neurological state of PD patients.

1.2.2 Prediction of the neurological state

Other studies consider speech signals recorded only in one session to predict the neurological state of PD patients. In [16] the authors proposed a methodology to predict the UPDRS-III score from the speech recordings of 82 subjects. The participants of the study were asked to perform three speech tasks including the sustained phonation of the vowel /a/, the repetition of three syllables (/pa/-/ta/-/ka/), and the reading of three standard passages. The

set of features extracted of the speech recordings include pitch, spectral entropy, thirteen cepstral coefficients, the number and duration of voiced and unvoiced frames, jitter, shimmer, HNR, and the ratio of energy in the first and second harmonics. The set of features were computed separately for each speech task and merged in a single feature vector. The UPDRS scores were predicted using two Support Vector Regressor (SVR)-based approaches: (1) epsilon-SVR and (2) nu-SVR. Additionally, different kernel functions were used to train the SVRs including polynomial, radial basis function, and sigmoid kernels. The authors reported that it is possible to predict the UPDRS-III with a MAE of 5.66 using an ε-SVR with a cubic polynomial kernel. Later in [17] the authors compared three regression techniques to predict the UPDRS scores including ridge regression, Lasso regression, and linear SVR. Speech recordings of 168 patients were collected in a single recording session. Additional to the features described before, the authors added information from features extracted with openSMILE [18]. Information of 21 HC were also included in the training process in order to analyze the influence of healthy speakers in the prediction of the disease severity according to the UPDRS-III. The authors report that the neurological state of the patients can be predicted with a mean absolute error of 5.5 considering only PD patients in the training process. On the other hand, in the INTERSPEECH 2015 Computational Paralinguistic Challenge (ComParE 2015) there was a Parkinson's Condition sub-challenge where the task of neurological state prediction of PD patients from speech was addressed [19]. Recordings of the 50 patients (25 male, 25 female) included in the PC-GITA database [20] were considered to form the train and development subsets. The test set included a total of 11 new patients recorded in non-controlled noise conditions, i.e., not using a sound-proof booth. A total of 42 speech tasks were considered. The neurological state of the patients was assessed by a neurologist expert according to the MDS-UPDRS-III subscale. The winners of the challenge reported a Spearman's correlation of 0.65 between the real MDS-UPDRS-III scores and the predicted values using Deep Rectifier Neural Networks (DRNN) and Gaussian processes Regression (GPR) [21].

In [22] it was presented a methodology to estimate the neurological state of PD patients. Speech recordings from Spanish, German, and Czech speakers were used to predict their UPDRS-III score using a linear ε-SVR. Four different speech tasks were considered for each database. For Spanish 50 PD patients (50 male, 50 female) were asked to pronounce 21 words, 6 sentences, one read text, and a monologue. For the German database a total of 88 patients (47 male, 41 female) were asked to pronounce 6 words, 5 sentences, one text, and a monologue. In the Czech database 20 male patients and the speech tasks considered include 7 words, 3 sentences, one text, and a monologue. The authors performed a model of articulation based on [23]. The transitions from unvoiced to voiced (onset) and from voiced

to unvoiced (offset) segments were modeled. The energy content of the onset and offset transitions was calculated in terms of 12 Mel-Frequency Cepstral Coefficients (MFCCs) and 22 Bark band energies (BBE). Additionally, speech intelligibility was evaluated using the Google Inc ® automatic speech recognition system. According to the authors the obtained results indicate that the neurological state of the patients, in terms of the MDS-UPDRS-III score, can be estimated with a Spearman's correlation of up to 0.74 when several speech tasks are modeled by the fusion of articulation and intelligibility measures.

Most of the studies in the literature are focused on predicting the neurological state of groups of PD patients from speech using different regression techniques. In general, the state of the disease is predicted using one recording session and there is no tracking of the disease progression over the time. For the monitoring of PD progression, the reviewed works are focused in detect changes in the features extracted from speech signals. However, the progression of the disease is not assessed individually for each patient. In this thesis is presented a methodology where the disease progression of PD patients was evaluated by adapting speaker models, i.e., it is proposed an approach for individual modeling of the disease progression that can be adapted to each patient individually. The proposed method is based on the Gaussian Mixture Model Universal Background Model (GMM-UBM) approach. The GMM-UBM systems are commonly used in speaker recognition due to their capability of representing a large class of sample distributions from which single-speaker models are obtained [24].
19-311. The speech signals are modeled following two different methods: (1) considering the energy content of the voiced and unvoiced segments and (2) the energy content of the transitions between voiced to unvoiced (offset) and unvoiced to voiced (onset) [23].

1.3 Hypothesis

The PD is a progressive disease that affects the motor capabilities of the patients, then it is possible to evaluate the disease progression from speech .

1.4 Contribution of this thesis

There are not many works in the literature regarding longitudinal analysis of PD patients from speech. Although there is interest in the scientific community to develop tools and methodologies for the continuous assessment of speech of PD patients, there are several limitations in most of the current studies. In general, the data collection process requires

some resources that are not always available, e.g., personal to record speech, expert clinicians to perform neurological examinations or speech assessments, ideal location to record speech, and others. In order to contribute to the studies on the monitoring of the PD, this thesis addressed the following aspects

- Longitudinal speech recordings: During the 2015 and the first semester of 2016, speech recordings from 27 PD patients were collected in three recording sessions. Further, 18 of the patients were assessed by a neurologist expert. Not all the patients are present on the three recordings sessions, however, the data collected during this work is complemented with the PC-GITA and the Multimodal databases created by the GITA research group in 2012 and 2014, respectively.
- Data collection with a portable device: The speech signals recorded in 2015 and 2016 were captured using a portable device based on the ODROID-U2. Additionally, the recordings were captured in non-controlled conditions and a classifier based on the Support Vector Machine was used to test the suitability of the recordings, yielding to satisfactory results [25].
- PD progression assessment from speech: The main contribution of this thesis is a methodology to assess the progression of the disease per speaker. The speech of the patients is modeled individually considering speech recordings captured in different sessions including PC-GITA and the Multimodal databases.

1.5 Structure of the thesis

Chapter 2 describes the methods used to assess the progression of PD patients from speech. The articulatory model described in Chapter 2 was first introduced in [23] and in this thesis is used to evaluate the suitability of the information that the voiced/unvoiced segments and the onset/offset transitions provided to assess speech of patients according to a standard neurological scale. Additionally, there is a description of the machine learning methods used to model speech and predict the neurological score of the patients.

Chapter 3 includes the information of the patients that participated in the recordings sessions and their stage of the disease according to the MDS-UPDRS. Additionally, there is information of the healthy control group and the set of patients used to train and test the models. Moreover, is presented the methodology implemented to assess the progression of PD over the time from speech recordings.

Chapter 4 contain details of the experiments performed on the described data. Two different approaches are considered in this Thesis to assess the disease progression from speech. The first one consists of a state-of-the-art regression technique used to predict the neurological state of the patients. The second approach is the proposed methodology which consists of individual speaker models used to track the PD progression over the time.

Chapter 5 includes a brief analysis of the results obtained in this work.

Chapter 6 summarizes the proposed approach.

Chapter 2

Analysis of speech of Parkinson's patients

2.1 Speech impairments in PD patients

Parkinson's disease (PD) is a neurological disorder caused by the degeneration of neurons within the brain, resulting in the progressive loss of dopamine in the substantia nigra of the midbrain [6]. The severity and progression of PD varies among patients, i.e., the progression of the disease and the symptoms suffered by some patients varies from person to person. The primary motor symptoms include resting tremor, slowness of movement , postural instability, rigidity and several dimensions of speech are affected including phonation, articulation, prosody, and intelligibility [26, 27]. Phonation impairments in PD patients include inadequate closing of the vocal fold and vocal fold bowing [28], which generates stability and periodicity problems in the vocal fold vibration. The articulation problems are mainly related with reduced amplitude and velocity of lip, tongue, and jaw movements [29], generating a reduced articulatory capability in PD patients to produce vowels [30] and to produce continuous speech. These deficits reduce the communication ability of PD patients and make their normal interaction with other people difficult.

2.1.1 Clinical diagnosis

There is no standard test to diagnose PD. Doctors relies on clinical history and physical examination to asses the patients. The disease severity is evaluated by neurologist experts by means of several tests. One of them is the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [11]. This is a perceptual scale used to assess motor and non-motor abilities of PD patients. The total MDS-UPDRS scale is divided into four

sections. The first section concerns the non-motor experiences of daily living such as cognitive impairment, depressed mood, and fatigue. The second section considers motor experiences of daily living including eating tasks, handwriting, and tremor, the third section concerns the motor examination including speech, finger taping, and gait, and the fourth section concerns motor complications such as time spend without medication.
In this thesis only the third section (MDS-UPDRS-III) is considered because it evaluates the motor capabilities of the patients. The section has a total of 33 items to evaluate different motor abilities but only one of them considers speech. However, the speech production process involves different organs to produce vocal and voiceless sounds e.g. vowels, consonants [31]. Thus, it makes sense to model motor capabilities from speech considering different aspects such as stability in the vocal folds vibration, energy content, articulatory capability, and others.
The speech of the patients is evaluated by clinicians considering the rating system presented in Table 2.1 [1]. During the examination, the patients are asked to talk about different subjects in order to assess volume of the voice, clarity in speech, modulation of words, among others.

Table 2.1 *Score system of the speech item from the MDS-UPDRS-III.*

Score	Rate	Definition
0	Normal	No speech problems
1	Slight	Loss of voice intensity or modulation
2	Mild	Some words are unclear
3	Moderate	Speech is difficult to understand
4	Severe	Speech is unintelligible

2.2 Articulatory model

The motor system to produce speech consists of several muscles and limbs which are involved in the phonation and articulation processes. The phonatory subsystem is in charge of producing voiced sounds by using the airflow from the lungs to make the vocal fold vibrate. The articulation subsystem consists of the set of muscles and limbs e.g., tongue, jaw, lips, velum, which are involved in the production of vowels and unvoiced sounds like consonants plosives and others. In the production of unvoiced sounds (UV) there is no vibration of the vocal fold and speech sounds are generated by turbulent airflow at a constriction in the vocal tract. During the production of the voiced segments the vibration of the vocal fold follows

[1] http://www.movementdisorders.org/MDS-Files1/PDFs/MDS-UPDRS-Rating-Scales/NewUPDRS7308final.pdf

four stages in one cycle: (1) closed, (2) opening, (3) open, and (4) closing. Figure 2.1 shows these four stages. There are several frequency and amplitude perturbation patterns which

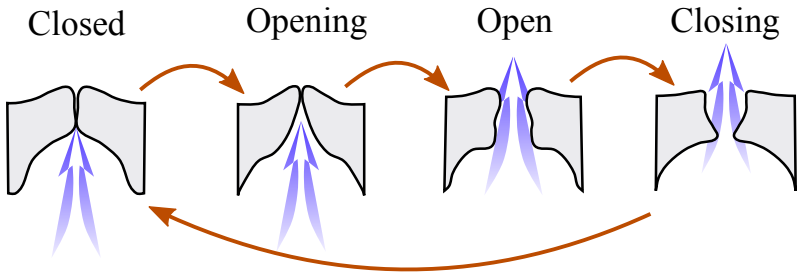

Fig. 2.1 *Vocal folds vibration pattern during voiced segments (Based on a figure found in [31])*

are observable during the production of vocal sounds (V). Those perturbations result from different factors such as the vocal fold asymmetry, involuntary movements at the larynx (neurogenic factors), and fluctuations of the airflow and subglottal pressure [31]. On the other hand, the unvoiced segments are produced by a total constriction at certain place in the vocal tract resulting on the interruption of the airflow. Unvoiced sounds are also produced by narrowing the air path producing turbulent airflow which creates the noise-like signals [32]. The method used in this work to identify voiced and unvoiced segments is based on the presence of the fundamental frequency of speech (pitch) in short-time frames as it was shown in [33]. Figure 2.2.A shows the pitch contour (red line) and two speech frames extracted from a voice recording. It can be observed that the voiced segments are quasi-periodic signals, while the unvoiced segments are noise-like signals. Additionally, the voiced to unvoiced

Figure 2.2.A. Voiced/unvoiced segments

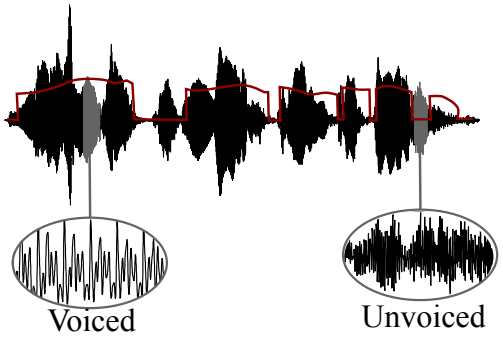

Figure 2.2.B. Onset/offset transitions

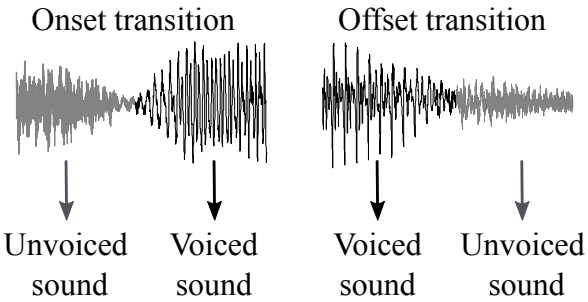

Fig. 2.2 *(A) Pitch contour (red line) and voiced/unvoiced short time windows extracted from a speech signal. (B) Onset and offset transitions frames*

transitions (offset transitions) and the unvoiced to voiced transitions (onset transitions) are considered to model the difficulties of the PD patients at the start and stop of the vibration

of the vocal folds (Figure 2.2.B) [23]. Onset and offset transitions are produced by the combinations of different sounds during continuous speech. Since neurological changes can be reflected in the speech production process, the information of the voiced/unvoiced segments as well as the onset/offset transitions is considered due to the different processes involved to generate speech. Figure 2.3 shows an spectrogram of the onset transition extracted from the speech recording of a PD patient and a healthy person. It can be observed that for the patient (Figure 2.3.A) the energy distribution is more disperse compared to the energy of the healthy person (Figure 2.3.B).

Figure 2.3.A. Onset from patient

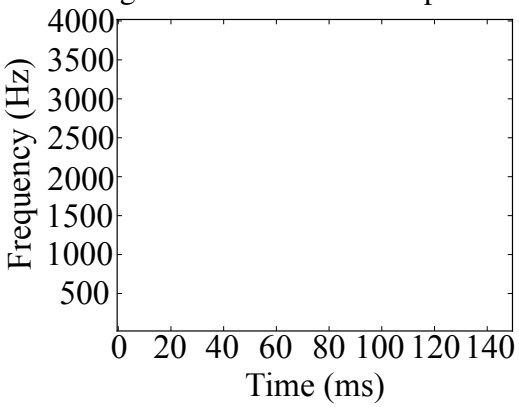

Figure 2.3.B. Onset from healthy person

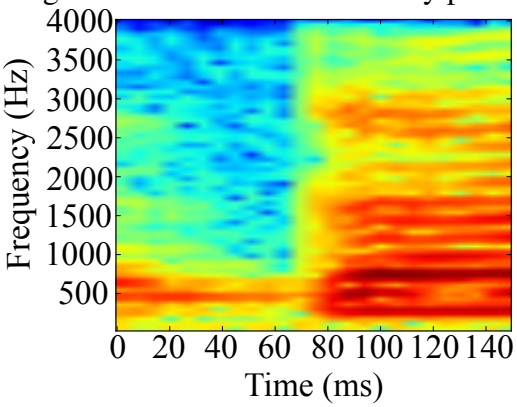

Fig. 2.3 *Spectrogram of the onset transition from (A) a PD patient, and (B) a healthy person.*

2.3 Speech modeling

The voiced/unvoiced segments and the onset/offset transitions are used to analyze speech impairments in PD patients considering phonation and articulation measures. Features extracted from the voiced segments are considered to model the temporal and amplitude variation of the vocal folds vibration. Articulation impairments are modeled considering spectral measures and the energy content of the unvoiced segments and the onset/offset transitions. The articulatory capability is also modeled extracting spectral features from the voiced segments.

2.3.1 Phonation measures

Jitter and shimmer:

Temporal and amplitude variations of the pitch period are defined as jitter and shimmer, respectively. Jitter is computed as in the Equation 2.1, where N is the length of the speech

signal in frames, M_p is the maximum pitch value, and $F_0(k)$ is the pitch value estimated in the k-th period.

$$\text{Jitter}(\%) = \frac{100}{N * M_p} \sum_{k=1}^{N} |F_0(k) - M_p| \tag{2.1}$$

Shimmer is estimated using the Equation 2.2, where $A(k)$ is the amplitude of the signal in the k-th period, and M_a is the maximum amplitude of the signal.

$$\text{Shimmer}(\%) = \frac{100}{N * M_a} \sum_{k=1}^{N} |A(k) - M_a| \tag{2.2}$$

Previous works have shown that the variation of the shimmer and jitter values are higher on patients compared to normal people, i.e., without any speech impairment or neurological disease [34]. Additionally, other works have shown that there is a significant change in shimmer when the speech of PD patients is analyzed in different recording sessions [13]. Figure 2.4 shows the difference between the fundamental frequency (pitch) contour extracted from the recordings of a PD patient and a healthy person. It can be observed that for a healthy person the pitch contour is more steady during the phonation compared to the pitch contour of the patient, resulting in higher values of jitter and shimmer.

Figure 2.4.A. F0 contour from patient

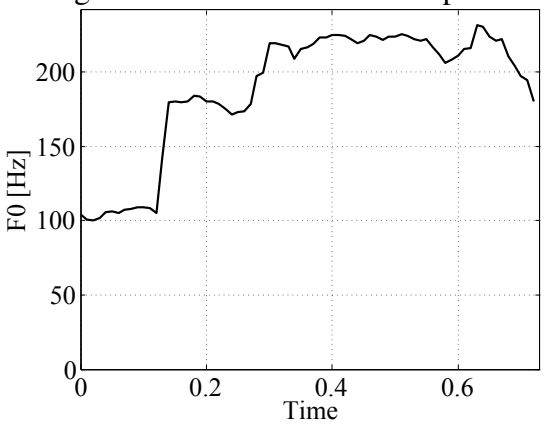

Figure 2.4.B. F0 contour from control

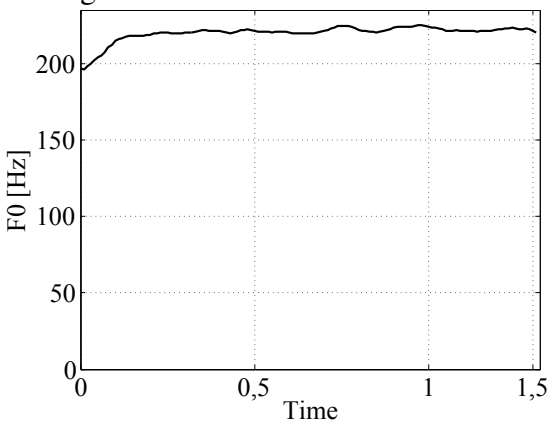

Fig. 2.4 *(A) Pitch contour estimated from the recording of a PD patient. (B) Pitch contour estimated from the recording of a healthy control*

2.3.2 Articulation measures

Mel Frequency Cepstral Coefficients:

These coefficients comprise a smoothed representation of the speech spectrum considering information of the scale of the human hearing. These features are widely used to model

articulatory problems in the vocal tract [35]. Commonly in speech processing 12 MFCCs are considered to analyze voice signals. The procedure to calculate the MFCCs consists of the following steps:

- The speech signal X is divided into segments $X = \{x_1, x_2, ..., x_n\}$.
- The discrete Fourier transformation is computed for each segment.
- The Mel filter-bank is applied. For speech signals captured at 44.1 kHz, the frequencies of the filter-bank ranges from 0 Hz up to 22.5 kHz. To convert from linear frequency to Mel scale the following equation is used.

$$M(f_{Hz}) = 1125 \ln(1 + f_{Hz}/700) \tag{2.3}$$

- The energy in each filter is added.
- The discrete cosine transform is applied upon the logarithm of the energy bands.

Bark band energies:
The Bark scale is a psycho-acoustic scale proposed by Zwicker in [36]. The scale ranges from 1 to 25. The division is performed according to the concept of the critical bands in the human auditory system. The conversion between frequency measured in Hertz and Bark scale is given by Equation 2.4. The Bark scale frequency bands are almost linear below 1 kHz, while from frequencies superior to 1 kHz the scale grows exponentially, which yields a perceptual filter-band ranging from 20 Hz up to 15.5 kHz. In this work the log-energy of the speech signal distributed in the 22 critical bands is calculated. The process to compute these energies consists of the following steps:

- Compute the short time Fourier transform (STFT) of the framed speech signal $X = \{x_1, x_2, ..., x_n\}$.
- Separate the corresponding spectrum in 25 frequency bands according to the Bark scale:

$$f_{Bark} = 13 \tan^{-1}\left(\frac{0.76 f_{Hz}}{1\,kHz}\right) + 3.5 \tan^{-1}\left(\left(\frac{f_{Hz}}{7.5\,kHz}\right)^2\right) \tag{2.4}$$

- Calculate the log-energy of each band [37].

2.4 Gaussian Mixture Model-Universal Background Model

The Gaussian Mixture Models (GMM)-based systems are capable of representing arbitrary probabilistic densities. GMMs are parametric probabilistic models represented as a weighted sum of M Gaussian densities (Figure 2.5). For a D-dimensional feature vector x a GMM is defined as:

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i p_i(x) \tag{2.5}$$

The Gaussian densities $p_i(x)$ are parameterized by the mixture weights ω_i, a D×1 mean vector μ_i, and a D×D covariance matrix Σ_i [38]. The parameters of the density models can be denoted as $\lambda = (\omega_i, \mu_i, \Sigma_i)$ and the Gaussian densities as

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\} \tag{2.6}$$

In speech processing GMMs are used to represent the distribution of feature vectors extracted

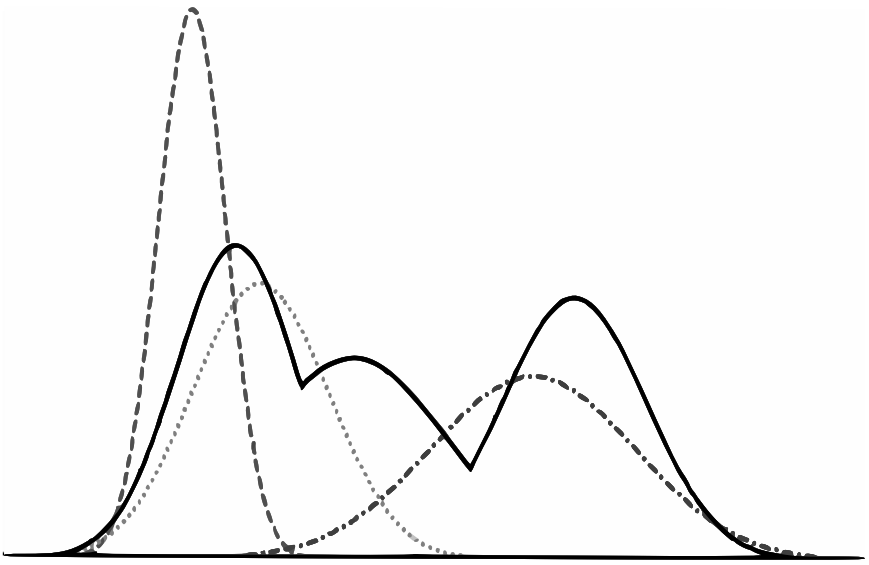

Fig. 2.5 *Graphical representation of a one-dimensional GMM. The solid black curve is the weighted sum of the Gaussian distributions represented by the gray dotted curves.*

from a single speaker or a group of speakers. If the GMM is trained using features extracted from a large sample of speakers, the resulting model is called Universal Background Model (UBM). Therefore, the UBM is trained to represent the entire space of possible speakers. For a given set of speakers, the conditional probability $p(X_{UBM}|\lambda)$ is known as the maximum likelihood function that better represents the population of speakers, where X_{UBM} are the set of feature vectors extracted from the group of speakers. The parameters λ of the maximum

likelihood function can be estimated using the Expectation Maximization (EM) algorithm. The EM approach is used to increase the likelihood of the UBM, i.e., for iterations k and $k+1$, $p(X|\lambda^{(k+1)}) > p(X|\lambda^{(k)})$. The steps of the EM algorithm are as follows:

- Initialize $\omega_k, \mu_k, \Sigma_k$. This is commonly achieved with a clustering algorithm such a K-means [39].
- Compute the new weights ω_{ik} with $1 \leq i \leq N$ and $1 \leq k \leq M$.

$$\omega_{ik} = \frac{p_k(x_i|\lambda_k)\omega_k}{\sum_{m=1}^{M} p_m(x_i|\lambda_m)\omega_m} \tag{2.7}$$

  Where N is the number of feature vectors extracted from the speakers, M is the number of Gaussian components in the GMM, $\omega_k = N_k/N$. N_k is the number of feature vectors contained in each Gaussian component.

- Compute the new mean vector μ_k

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \omega_{ik} x_i \tag{2.8}$$

- Compute the new covariance matrix Σ_k

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \omega_{ik}(x_i - \mu_k)(x_i - \mu_k)^T \tag{2.9}$$

Then ω_{ik} is computed again and the procedure is repeated until the convergence. In a GMM-UBM system the single speaker model is derived from the population of speakers by adapting the parameters of the UBM using the training data from the speaker to be modeled. There are different approaches used to obtain the speaker model. One method is the Maximum A Posteriori (MAP) adaptation which consist of a two step estimation process. In the first step the training vectors from the speaker to be modeled are aligned into the UBM mixture components. That is, given a UBM and the training vectors from the speaker $X = \{x_1, x_2, ..., x_T\}$, for mixture i in the UBM, we compute

$$Pr(i|x_t) = \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_j p_j(x_t)} \tag{2.10}$$

Then $Pr(i|x_t)$ and x_t are used to compute the sufficient statistics of the weight, mean, and variance

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \tag{2.11}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t \tag{2.12}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t^2 \tag{2.13}$$

The second step of the adaptation process consist of using these sufficient statistics to update the parameters of the UBM for the mixture i using the following equations

$$\hat{\omega}_i = [\alpha_i^{\omega} n_i / T + (1 - \alpha_i^{\omega}) \omega_i] \gamma \tag{2.14}$$

$$\hat{\mu}_i = \alpha_i^{m} E_i(x) + (1 - \alpha_i^{\omega}) \mu_i \tag{2.15}$$

$$\hat{\sigma}_i^2 = \alpha_i^{v} E_i(x^2) + (1 - \alpha_i^{v})(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2 \tag{2.16}$$

Where $\{\alpha_i^{\omega}, \alpha_i^{m}, \alpha_i^{v}\}$ are the adaptation coefficients used to control the balance between the old and new estimates for the weights, means, and variances, respectively. The scale factor, γ, is computed to ensure that the weights sum to unity. T is the number of feature vectors extracted from the speaker [38]. The resulting adapted model can be used to assess the progression of the disease from the patients, considering the changes with respect to the UBM. Further details of this procedure are described in Chapter 3.

2.5 Support Vector Regression

The goal of the Support Vector Regression (SVR) proposed by Vapnik [40] is to find a function $f(x)$ that has at most ε deviation from the original targets $y_i \in R$. The main idea is to minimize the prediction error $|y - f(x)|$ using the same principles of the support vector for classification: obtain a hyperplane which maximizes the margin considering that part of the error is tolerated. In this case, the prediction error can be above or below the target value and the margin is described by $f(x_i) - y_i > \varepsilon$ and $y_i - f(x_i) > \varepsilon$ [23]. In other words, the prediction error is tolerated as long as they are less than ε. In the case of linear regression, the prediction function can be estimate as:

$$f(x) = \langle w, x \rangle + b \tag{2.17}$$

With $\mathbf{w} \in \mathbf{X}$, $b \in R$, where X represent the space of input patterns $\{x_1, x_2, ..., x_d\}$. The optimization problem can be formulated as

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M} |y_i - f(x_i)|_\varepsilon \tag{2.18}$$

Where C is a penalty parameter which determines the amount of trade-off between the flatness of $f(x)$ and the amount of error larger than ε which is tolerated [41]. To formulated Equation 2.18 as a constrained optimization problem, the slacks variables ξ and ξ^* are introduced. Assigning ξ to $f(x_i) - y_i > \varepsilon$ and ξ^* to $y_i - f(x_i) > \varepsilon$. The primal objective function can be expressed as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - \langle w, x_i\rangle + b \leq \varepsilon + \xi_i \\ & \langle w, x_i\rangle + b - y_i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{aligned} \tag{2.19}$$

Figure 2.6 illustrates the situation. Only the points outside the shaded region (ε-insensitive tube) contribute to the cost insofar, as the prediction errors are penalized in a linear fashion. One way to solve the optimization problem presented in 2.19 is in its dual formulation

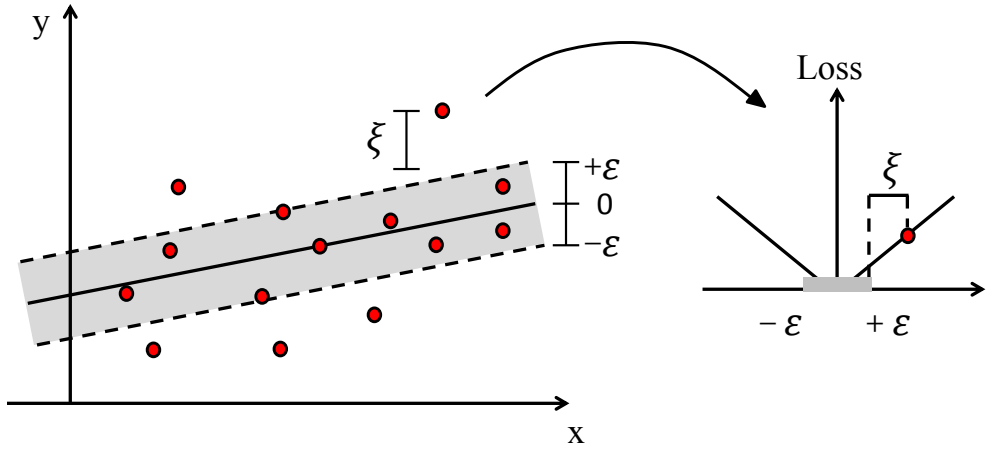

Fig. 2.6 *Linear Support Vector Regressor (Based on a figure found in [23]).*

using Lagrange multipliers. The main idea in the dual formulation is to construct the Lagrange function from the primal function (objective function) and the constrains are obtained introducing a dual set of variables. The Lagrange function of the primal problem is

expressed as

$$\begin{aligned} L = \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*) - \sum_{i=1}^{M}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ & - \sum_{i=1}^{M}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ & - \sum_{i=1}^{M}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) \end{aligned} \tag{2.20}$$

Where $\{\alpha_i, \alpha_i^*, \eta_i, \text{ and } \eta_i^*\} \geq 0$. The primal variables (w, b, ξ_i, ξ_i^*) have to vanish for optimality. This is accomplished estimating the partial derivatives of L with respect the primal variables:

$$\begin{aligned} \partial_b L = \quad & \sum_{i=1}^{M}(\alpha_i^* - \alpha_i) = 0 \\ \partial_w L = \quad & w - \sum_{i=1}^{M}(\alpha_i - \alpha_i^*)x_i = 0 \\ \partial_{\xi_i} L = \quad & C - \alpha_i - \eta_i = 0 \\ \partial_{\xi_i^*} L = \quad & C - \alpha_i^* - \eta_i^* = 0 \end{aligned} \tag{2.21}$$

Substituting the previous results in 2.20, the dual optimization problems is expressed as

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle - \varepsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M} y_i(\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{M}(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i, \alpha_j^* \in [0, C] \end{aligned} \tag{2.22}$$

From the partial derivatives we have that $w = \sum_{i=1}^{M}(\alpha_i - \alpha_i^*)x_i$. The regression function can be rewritten as

$$f(x) = \sum_{i=1}^{M}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \tag{2.23}$$

Now the weights w can be described as a linear combination of the training patterns x_i. In order to compute b the Karush-Kuhn-Tucker (KKT) conditions are verified. The KKT conditions state that at the point of the solution the product between dual variables and constrains has to vanish

$$\begin{aligned} & \alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) = 0 \\ & \alpha_i^*(\varepsilon + \xi_i^* - y_i + \langle w, x_i\rangle - b) = 0 \end{aligned} \tag{2.24}$$

and

$$\begin{aligned}(C-\alpha_i)\xi_i &= 0\\ (C-\alpha_i^*)\xi_i^* &= 0\end{aligned} \tag{2.25}$$

From the previous results can be concluded that only samples (x_i, y_i) with corresponding $\alpha_i = C$ and $\alpha_i^* = C$ lie outside the ε-insensitive tube. Additionally, there is never a set of dual variables α_i, α_i^* which are both simultaneously nonzero ($\alpha_i \alpha_i^* = 0$). From this we obtain that

$$b = y_i - \langle w, x_i \rangle + \varepsilon \text{ for } 0 < \alpha_i < C \tag{2.26}$$

and

$$b = y_i - \langle w, x_i \rangle - \varepsilon \text{ for } 0 < \alpha_i^* < C \tag{2.27}$$

Finally, from 2.24 it follows that for all samples inside the ε-insensitive tube the Lagrange multipliers α_i, α_i^* vanish for $|f(x_i) - y_i| < \varepsilon$.

2.6 Probabilistic distance measures

Some of the probabilistic distances used in this work were proposed as methods to provided bounds for the probability of error [42]. It is not the purpose of this thesis to review the full procedure on how to formulate the distance measures. However, the basic concepts are reviewed here in order to provide a better understanding of the methods. A more complete analysis can be found on the books of Richard D. [42] and Devijver and Kittler [43].
To derive the distance measure between density distributions consider the case of a bi-class $(c1, c_2)$ classifier based on the Bayesian decision rule. If there is an observation x for which $P(c_1|x) \geq P(c_2|x)$, then we assigned x to the class c_1. Similarly if $P(c_1|x) \leq P(c_2|x)$, then x is assigned to c_2. The probability of error is calculated whenever an observation pattern is assigned to one class, that is $P(error|x) = P(c_1|x)$ if x is assigned to c_2 and $P(error|x) = P(c_2|x)$ otherwise. The probability of error can be minimize if x is assigned to its true class. However, there are other observations that will not have the same value of x, but still belongs to the same class. In this case, the average probability error is expressed as

$$P(error) = \int_{\infty}^{-\infty} P(error|x)p(x)dx \tag{2.28}$$

It follows that, whenever a pattern x is observed, then

$$P(error|x) = \min\left[P(c_1|x), P(c_2|x)\right] \tag{2.29}$$

Where

$$P(c_i|x) = \frac{p(x|c_i)P(c_1)}{p(x)}, \text{ with } i = 1,2. \tag{2.30}$$

Chernoff bound:
To estimate the Chernoff distance, the following inequality is considered.

$$\min[a,b] \leq a^{\beta} b^{1-\beta}, \text{ for } a,b \geq 0 \text{ and } 0 \leq \beta \leq 1 \tag{2.31}$$

Assuming $a \geq b$, the previous inequality is re-written as $b \leq a^{\beta} b^{1-\beta} = \left(\frac{a}{b}\right)^{\beta} b$. Using Equations 2.30 and 2.28 and appltying the inequality to 2.29, we get the bound expressed as

$$P(error) \leq P^{\beta}(c_1)P^{1-\beta}(c_2)| \int p^{\beta}(x|c_1)p^{1-\beta}(x|c_2)dx \text{ for } 0 \leq \beta \leq 1 \tag{2.32}$$

For the case of conditional probabilities with Gaussian distribution, the integral in Equation 2.33 can be evaluated analytically to obtain:

$$\int p^{\beta}(x|c_1)p^{1-\beta}(x|c_2)dx = \mathrm{e}^{-f(\beta)} \tag{2.33}$$

The Chernoff distance is computed with $d_{Che} = -\log(\mathrm{e}^{-f(\beta)}) = f(\beta)$. For Gaussian distributions, $d_{Che} = f(\beta)$ is expressed as

$$d_{Che} = \frac{1}{2}\beta(1-\beta)(\mu_2-\mu_1)^T[\Sigma_{\beta}]^{-1}(\mu_2-\mu_1) + \frac{1}{2}\log\left(\frac{|\Sigma_{\beta}|}{|\Sigma_1|^{1-\beta}|\Sigma_2|^{\beta}}\right) \tag{2.34}$$

Where $\Sigma_{\beta} = 1(1-\beta)\Sigma_1 + \beta\Sigma_2$.

Bhattacharyya distance:
Similarly to the Chernoff distance, the Bhattacharyya distance can be obtained evaluating the probability error function. Considering that the Bhattacharyya distance is a particular case of the Chernoff distance for $\beta = 0.5$, from Equation 2.34 we get that:

$$d_{Bha} = \frac{1}{4}(\mu_2-\mu_1)^T[\Sigma_1+\Sigma_2]^{-1}(\mu_2-\mu_1) + \frac{1}{2}\log\left(\frac{|\Sigma_1+\Sigma_2|}{2(|\Sigma_1||\Sigma_2|^{\frac{1}{2}})}\right) \tag{2.35}$$

Is the Bhattacharyya distance between probability distributions. For GMMs, the Bhathacharyya distance can be estimated as:

$$d_{Bha} = \frac{1}{8}\sum_{i=1}^{M}\left\{(\widehat{\mu}_i-\mu_i)^T\left[\frac{\widehat{\Sigma}_i+\Sigma_i}{2}\right]^{-1}(\widehat{\mu}_i-\mu_i)\right\}+\frac{1}{2}\sum_{i=1}^{M}\left[\ln\frac{\left|\frac{\widehat{\Sigma}_i+\Sigma_i}{2}\right|}{\sqrt{\left|\widehat{\Sigma}_i\right||\Sigma_i|}}\right]-\omega_{\text{Bha}} \quad (2.36)$$

Where $\widehat{\mu}_i$ and $\widehat{\Sigma}_i$ are the mean vector and covariance matrix from the UBM, μ_i and Σ_i are the mean vector and covariance matrix from the speaker model, and $\omega_{Bha} = \frac{1}{2}\sum_{i=1}^{M}\ln(\widehat{\omega}_i\omega_i)$ is the mixture weight measure [44].

Kullback-Leibler divergence:
This measure can be used to represent the difference between distribution. The Kullback-Leibler (KL) divergence between two probability distributions is expressed as

$$KL(p_1\|p_2) = \int p_1(x)\log\left(\frac{p_1(x)}{p_2(x)}\right)dx \quad (2.37)$$

For GMMs, the KL divergence is expressed as

$$KL = \sum_{i=1}^{M}\frac{\omega_i}{2}\left\{\log\left(\frac{|\widehat{\Sigma_i}|}{|\Sigma_i|}\right)+\text{Trace}((\widehat{\Sigma}_i)^{-1}\Sigma_i)+(\widehat{\mu}_i-\mu_i)^T(\widehat{\Sigma}_i)^{-1}(\widehat{\mu}_i-\mu_i)-D\right\} \quad (2.38)$$

However this measure is neither positive definite nor symmetric [44]. To computed the distance between distributions a symmetric version of the KL divergence is used: $d_{KL} = KL_1(p_1\|p_2)+KL_2(p_2\|p_1)$.

2.7 Entropy measures

There are different ways to compute the changes of the individual speaker models respect to the UBM. A different approach to the probabilistic distance measures is based on the amount of information provided by the adapted models. In [45] Claude Shannon introduced a measure suitable to quantified the amount of information in a signal. This measure was called entropy. In this thesis, the concept of entropy is used to measure the amount of information present in the individual speaker models in order to detect changes in speech.

Shannon entropy:
Let $P = (p_1, p_2, ..., p_k)$ be a finite discrete probability distribution. The entropy of the dis-

tribution P is described as the amount of uncertainty of possible results which have the probabilities $p_1, p_2, p_3, ..., p_n$. The Shannon entropy of the distribution P is defined as

$$H_{Shann} = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} \tag{2.39}$$

For the adapted model, obtained from a UBM, the probabilities p_k are estimated using Equation 2.5, were x and $\lambda_{Speaker}$ are the feature vector and the parameters from the speaker model.

Rényi entropy:
This measure is presented by Álfred Rényi in [46] and is introduced as a generalization of the Shannon entropy. Rényi's entropy measure is defined as

$$H_{Renyi} = \frac{1}{1-\alpha} \log_2 \sum_{k=1}^{n} p_k^{\alpha} \tag{2.40}$$

Were α is a non-negative parameter with $\alpha \neq 1$ and $\alpha > 0$.

Relative entropy:
Since the UBM is formed with a higher number of samples than the adapted model, it is not possible to measure the relative entropy between the two models. However, a similar approach was already introduced in Section 2.6. The relative entropy of two probabilistic distributions (P, Q) is defined as

$$H_{Rel}(P \| Q) = \sum_{k=1}^{n} p_k \log_2 \frac{p_k}{q_k} \tag{2.41}$$

Note that this measure has the same form of the Equation 2.37. In fact, the relative entropy is also known as the Kullback-Leibler divergence. Instead of using relative entropy measures, only the probabilistic distances described in Section 2.6 are used to measure the changes of the adapted models respect to the UBM.

Chapter 3

Methodology

Figure 3.1 summarizes the stages of the methodology followed in this study. In the first stage one patient is selected to be modeled and the remaining speakers are used for training. Then, the voiced and unvoiced segments and the onset and offset transitions are extracted from the speech recordings to perform the feature estimation process. The features from the training set are used to create the universal model. The set of features from the selected patient are used to obtain an individual model adapted from the universal model. The user model per speaker consists of three single models, one per recording session. Finally, the disease progression is evaluated calculating the distance between the universal model and the user model. Further details are provided in the following subsections. The disease

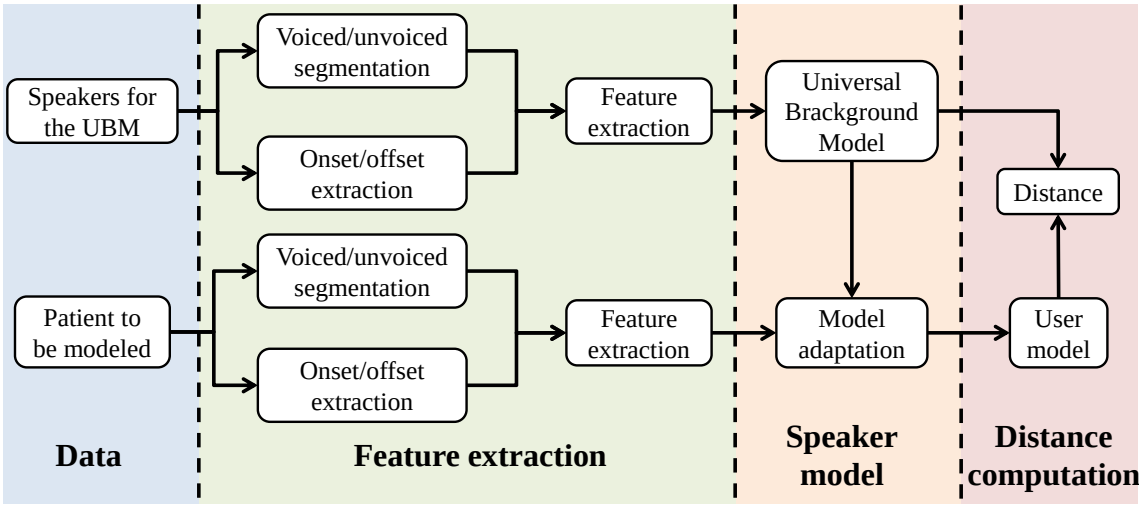

Fig. 3.1 *General methodology used to built speaker models*

progression is evaluated considering the articulatory model presented in [23]. The main reason is that according to previous work, the best results are always obtained when the

articulatory approach is considered to detect PD and to predict the neurological state of the patients.
[22, 25, 33, 47].

3.1 Data

Speech recordings from 64 PD patients (34 male and 30 female) were collected in three recording sessions from 2012 to 2015. Most of the patients participated at least in two of the three recording sessions. A subset of seven patients participated in all recording sessions, thus their speech samples are considered here to test the speaker models. A professional audio setting was used for the first two sessions, and the remaining sessions were recorded in non-controlled acoustic conditions, i.e., not using a sound-proof both and a professional audio setting, using the device presented in [25]. The speech recordings were captured with a sampling frequency of 44.1 kHz with a resolution of 16 bits.
Additionally, in [25] the suitability of the speech recordings for detecting PD was evaluated considering a Support Vector Classifier-based method, leading to satisfactory results. Thus, the speech recordings collected with the device were considered for the speaker models.
All of the patients were evaluated by a neurologist expert according to the MDS-UPDRS-III [11] at the moment of the recordings. The MDS-UPDRS-III scores of the seven patients used to test the speaker models are provided in Table 3.1.

Table 3.1 *Distribution of patients recorded in all sessions.* ***Session*** **i** *($i \in \{1,2,3\}$): MDS-UPDRS-III scores obtained on each recording session.*

Patient	**Gender**	**Age**	**MDS-UPDRS-III**		
			Session 1	**Session 2**	**Session 3**
P1	M	64	28	19	13
P2	M	59	6	8	24
P3	M	68	14	25	7
P4	F	55	29	26	26
P5	F	57	41	35	33
P6	F	51	38	49	44
P7	F	55	43	10	19

Additionally, a set of healthy control (HC) speakers is formed with recordings from 64 persons (32 male, 32 female). None of the participants in the HC group has a history of symptoms related to PD or any other kind of movement disorder. Each subject in the HC group was recorded once. All of the participants of the test followed the set of speech tasks

Table 3.2 *Distribution of patients and healthy controls recorded in the remaining sessions.* ***Session*** **i** *($i \in \{1,2,3\}$): MDS-UPDRS-III scores obtained on each recording session.*

Male PD patients				**Male HC**	**Female PD patients**				**Female HC**
Age	**MDS-UPDRS-III**			**Age**	**Age**	**MDS-UPDRS-III**			**Age**
	Session 1	**Session 2**	**Session 3**			**Session 1**	**Session 2**	**Session 3**	
65	32	43	-	67	72	19	19	-	63
60	44	-	-	67	75	52	-	-	75
81	50	28	-	55	66	28	18	-	65
57	20	-	-	55	55	30	38	-	60
77	92	-	-	56	60	29	-	-	57
50	53	-	-	63	57	61	35	-	63
75	13	-	-	42	66	28	-	-	73
75	75	-	-	65	55	30	38	-	55
56	30	-	-	86	62	42	19	-	68
50	19	-	-	63	61	21	20	-	62
74	40	-	-	76	69	19	-	-	61
48	9	-	-	61	59	40	33	-	65
68	67	-	-	51	51	23	-	-	63
54	15	-	-	62	65	54	-	-	55
33	51	-	-	67	59	71	-	-	63
69	40	-	-	68	64	40	-	93	58
67	28	-	-	54	49	53	-	-	62
47	33	-	-	67	73	38	64	-	61
65	53	-	-	71	58	57	47	-	64
64	45	-	-	50	70	23	20	-	76
68	65	-	-	62	54	30	14	-	61
45	21	-	-	68	71	-	27	39	57
67	-	58	63	64	61	-	36	27	50
70	-	64	37	31	53	-	31	-	49
61	-	82	-	42	65	-	46	-	50
70	-	40	-	55	75	-	-	22	50
66	-	12	-	60	-	-	-	-	52
48	-	13	-	61	-	-	-	-	54
78	-	61	-	62	-	-	-	-	54
60	-	-	40	75	-	-	-	-	59
64	-	-	45	78	-	-	-	-	70
-	-	-	-	78	-	-	-	-	78

presented in [20]. For this thesis three speech tasks were considered: the repetition of the word /pa-ta-ka/, a monologue, and the reading of a phonetically balanced text which contains 36 words:

Ayer fui al médico. ¿Qué le pasa? Me preguntó. Yo le dije: Ay doctor! Donde pongo el dedo me duele. ¿Tiene la uña rota?. Sí. Pues ya sabemos qué es. Deje su cheque a la salida.

The MDS-UPDRS-III scores of the remaining 57 patients used to train the universal background models and the age of the healthy controls are provided in Table 3.2.

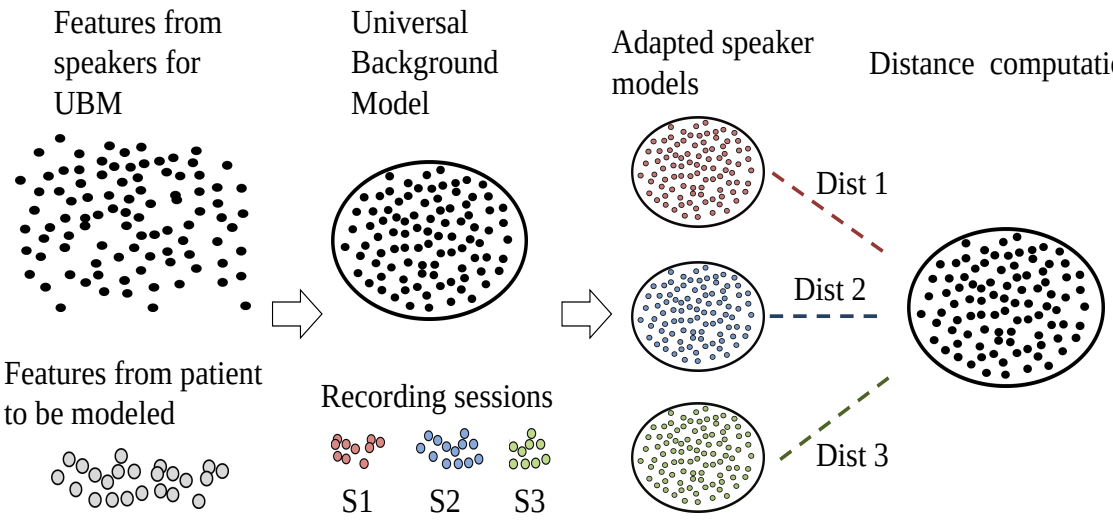

Fig. 3.2 *User modeling methodology.*

3.2 Feature extraction

Voiced/unvoiced segments and onset/offset transition were extracted as in [33] and [47] respectively. The decision whether the segment is voiced or unvoiced was made using Praat [48]. Hamming windowing with 20 ms length and a time-shift of 10 ms is applied. The onset/offset transitions are detected using frame windows of 80 ms length. Then, the features are extracted. For voiced frames the set of features includes jitter, shimmer, and 12 Mel–Frequency Cepstral Coefficients (MFCCs), forming a 14-dimensional feature vector. The unvoiced frames and the onset and offset transitions are modeled computing 12 MFCCs and the log energy of the signal distributed in 22 Bark bands, forming a 37-dimensional feature vector.

3.3 Speaker model

The user models-based approach are generated from speech recordings for the patients described in Table 3.1. One of the 64 patients from the database is extracted to be modeled. These individual speaker models are adapted from a background model which is considered as the baseline to assess the disease progression according to its distance to the adapted speaker. The trained model is adapted using the Maximum A Posteriori (MAP) approach. To observe the influence of a healthy control group in the performance of the user models, three different background models are considered: (1) with recordings of PD patients, (2) with healthy controls, and (3) with the combination of both groups of speakers. The models are built with

the features estimated from the voiced/unvoiced segments and the onset/offset transitions. The speaker models consists of three single models, one per recording session. The disease progression is evaluated by calculating the distance between the background model and the speaker model. Finally, the correlation between the distance measures estimated for each recording session and the three neurological scores is calculated. The details of the procedure are depicted in Figure 3.2.

3.4 Distance computation

The speech of PD patients can be assessed using the individual speaker models obtained from the GMM-UBM approach. The resulting models are based on probabilistic representations of the articulatory model described in Section 2.2. Hence, the probabilistic distance measures described in Section 2.6 are used to detect changes in speech of PD patients. Parkinson's is a progressive disease, which means that the severity of the symptoms get worse over the time. These changes in the state of the disease may be reflected in the neurological examination performed by the expert. Considering that the patients were assessed according to the MDS-UPDRS-III scale, the neurological score is expected to increase through sessions.

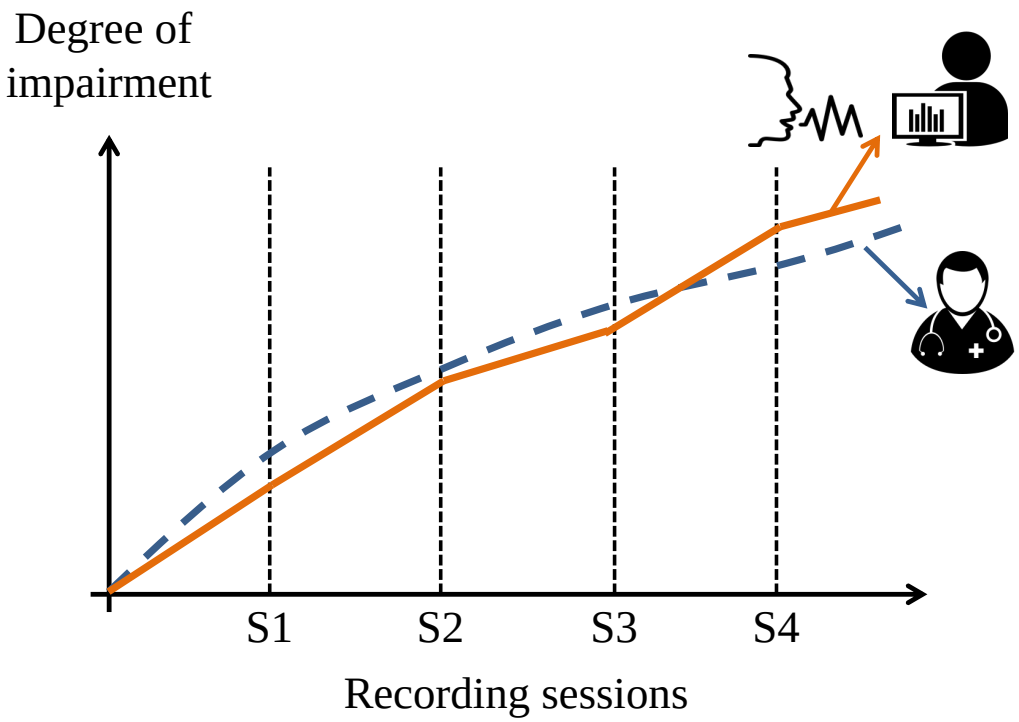

Fig. 3.3 *Graphical representation of the progression of PD.*

According to the studies reviewed in the literature, the speech of the PD patients is impaired and such an impairment is progressive [13]. The goal of the speaker models is to identify the changes in the speech of the patients according to the probabilistic distance between the background model and the speaker model. Since the distances are estimated with speech recordings captured at the same time of the neurological examination, it is expected that the trend of these distances follows the trend of the MDS-UPDRS-III. Figure 3.3 shows a graphical representation of the described situation. The dotted blue curve represents the trend of the disease progression for a patient assessed in different sessions (S1, S2, S3, and S4). According to the progression of the disease, the MDS-UPDRS-III score may show an increment between sessions. If these neurological changes are reflected in the speech of the patients, then the estimated distances should coincide with the trend of the changes in the neurological scores (orange continuous line).

Chapter 4

Experiments and results

Two different approaches are used to assess the PD progression in speech. The first approach is based on a linear ε-SVR which is used to validate the features extracted from the recordings of the PD patients. The second approach consist of a GMM-UBM-based approach which is used to obtain individual speaker models and to evaluate the progression of PD over the time. Each model was train considering three different speech tasks: a read text, a monologue, and the repetition of the word /pa-ta-ka/. Additionally, the features from the voiced/unvoiced segments and the onset/offset transitions were considered separately to train the models. The performance of the models is evaluated estimating the Pearson's correlation coefficient r between the predicted scores and the real MDS-UPDRS-III scores (in the case of the ε-SVR) and the Pearson's correlation coefficient between the estimated distances and the MDS-UPDRS-III scores (in the case of the GMM-UBM). It is important to note that in the beginning, the Spearman's correlation was also considered; however, since this measure works with ranked variables, the obtained results were biased due to the low amount of data (3 samples of the predicted scores, the estimated distances, and the MDS-UPDRS-III scores).

4.1 Regression model

The suitability of the features to predict the MDS-UPDRS-III scores is evaluated using a ε-SVR. The training sets are formed with features extracted from the PD patients presented in Table 3.2. The optimization is performed following a leave-one-speaker-out cross-validation (LOSO-CV) strategy. The parameters of the regressor C and ε are optimized in a grid search with $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ and $\varepsilon \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 20\}$. The ε-SVR is tested with the patients described in Table 3.1. The performance is evaluated predicting the neurological score of the patients using the features extracted from the three recording sessions.

4.1.1 Validation of voiced and unvoiced features

The first experiment consist of the prediction of the neurological scores of the PD patients considering the features extracted from the voiced/unvoiced segments. Table 4.1 shows the Pearson's correlation obtained per patient. In general, the highest average correlations were obtained for the unvoiced segments (read text: $r = 0.34$; monologue: $r = 0.06$; /pa-ta-ka/: $r = 0.31$) compared to the voiced segments (read text: $r = 0.03$; monologue: $r = -0.04$; /pa-ta-ka/: $r = -0.22$). The best results were obtained for the dedicated speech tasks, i.e, the read text and the repetition of /pa-ta-ka/.

Table 4.1 *Pearson's correlation between the predicted score and the real MDS-UPDRS-III.* ***Seg****: Voiced/unvoiced segments.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between the predicted score and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

Training set considering features from voiced/unvoiced segments									
Speech Task	**Seg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	V	0.70	-0.95	0.82	0.38	0.68	-0.99	-0.42	0.03
	UV	0.99	-0.99	0.92	0.56	0.39	-0.07	0.57	0.34
Monologue	V	0.84	-0.97	0.98	-0.09	0.12	-0.86	-0.33	-0.04
	UV	0.97	-0.99	0.60	0.46	-0.07	-0.95	0.37	0.06
/Pa-ta-ka/	V	0.11	-0.99	0.04	0.95	-0.45	-0.99	-0.23	-0.22
	UV	0.59	-0.74	0.77	0.98	0.60	-0.70	0.69	0.31

Note that the correlation coefficients estimated for P1, P3, P4, and P7 are improved for most of the speech tasks when the unvoiced segments are considered, resulting in a higher performance compared to the voiced segments. These results are mainly affected by the performance obtained when P7 is modeled. For this patient there are strong variations in the correlation values obtained for each task. This can be explained considering that for this patient the MDS-UPDRS-III scores show a strong change in the state of the disease from session 1 (MDS-UPDRS-III: 43) to session 2 (MDS-UPDRS-III: 10) which makes it harder to model using this approach.

4.1.2 Validation of onset and offset features

The same procedure used for the voiced/unvoiced segments is used to predict the neurological state of the patients considering features extracted from the onset and offset transitions. Table 4.2 shows the obtained results. Note that, in general, there is an improvement in the average correlation obtained for the read text (offset: $r = 0.45$) and the monologue (onset: $r = 0.27$). For /pa-ta-ka/ the results obtained were similar in both onset ($r = 0.25$) and offset ($r = 0.29$)

transitions.

Table 4.2 *Pearson's correlation between the predicted score and the real MDS-UPDRS-III.* ***Tran****: Onset/offset transitions.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between the predicted score and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

Training set considering features from onset/offset transitions									
Speech Task	**Tran**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	Onset	0.84	-0.99	0.71	0.72	0.19	-0.17	-0.19	0.16
	Offset	0.98	-0.98	0.98	0.99	0.50	0.47	0.23	0.45
Monologue	Onset	0.95	-0.99	0.79	0.21	0.99	-0.41	0.34	0.27
	Offset	0.50	-0.77	0.77	0.77	0.04	-0.63	-0.12	0.08
/Pa-ta-ka/	Onset	0.66	-0.97	0.92	0.82	0.90	-0.93	0.33	0.25
	Offset	0.96	-0.98	0.60	0.76	0.99	-0.72	0.40	0.29

Note that for P1, P2, and P3, the overall performance of the models shown in Table 4.2 are similar to the results presented in Table 4.1. The consistency in the results indicate that for these patients the articulatory model and the features used are suitable to model the progression of the disease. Although for P2 a negative correlation is obtained in all cases, the performance of the models (average $r = -0.94 \pm 0.09$) indicate that it is possible to evaluate its disease progression from speech. Additionally, for P4 and P5 the correlations improved when the features from the onset and offset transitions are considered.

4.2 Individual speaker/patient models

Three different versions of the universal background model are considered: (1) with recordings of a total of 63 PD patients, (2) with 64 healthy speakers, and (3) with both groups of speakers. The aim of these experiments was to determine if the information from the healthy speakers is suitable to improve the performance of the user models. The number of Gaussians used to train the UBM ranges from 2 to 512 in steps of 2^n ($n \in \{1,2,3...,9\}$).

4.2.1 Experiments with PD patients

Table 4.3 shows the Pearson's correlation between the Bhattacharyya distances and the MDS-UPDRS-III scores for the three speech tasks when features from the voiced and unvoiced segments were considered for training. It can be observed that the highest correlations per trained models are obtained for the unvoiced segments with respect to the voiced segments considering the read text ($r = 0.45$) and monologue ($r = 0.29$) speech tasks. For /pa-ta-ka/

similar results were obtained considering features from the voiced ($r = 0.38$) and unvoiced segments ($r = 0.32$). For the case of the features from the onset/offset transitions, the

Table 4.3 *Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Seg****: Voiced/unvoiced segments.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

UBM trained considering features from voiced/unvoiced segments										
Speech Task	**M**	**Seg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	128	V	0.84	-0.92	0.77	0.86	0.99	-0.54	0.82	0.40
	256	UV	0.96	-0.45	0.48	0.99	0.93	0.84	-0.62	0.45
Monologue	256	V	0.69	0.99	-0.21	-0.57	-0.89	0.45	-0.99	-0.08
	16	UV	0.52	0.83	0.70	-0.89	0.69	0.90	-0.75	0.29
/Pa-ta-ka/	4	V	-0.26	0.74	0.24	-0.02	0.92	0.91	0.12	0.38
	4	UV	0.99	-0.66	-0.85	0.29	0.99	0.98	0.53	0.32

results are show in Table 4.4. In this case the highest correlations per trained models are obtained for the offset transitions with respect to the onset transitions considering the read text ($r = 0.54$) and monologue ($r = 0.59$) speech tasks. For /pa-ta-ka/ the best result was obtained considering features from the onset transitions ($r = 0.26$). Considering the results from Tables 4.3 and 4.4, it can be observe that the highest correlations were obtained considering the features from the onset/offset transitions for training. In particular the best results were obtained considering features from the offset transitions for the read text and the monologue. Although the overall best result was obtained for the monologue, the read text had the most consistent performance in both experiments.

Table 4.4 *Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Tran****: Onset/offset transitions.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

UBM trained considering features from onset/offset transitions										
Speech Task	**M**	**Tran**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	32	Onset	0.52	0.92	0.89	0.12	0.99	-0.48	-0.49	0.35
	128	Offset	0.99	0.92	0.81	-0.25	0.88	0.95	-0.55	0.54
Monologue	32	Onset	0.38	-0.55	0.92	0.47	0.95	0.01	0.80	0.42
	32	Offset	0.19	0.74	0.97	0.65	0.91	0.34	0.34	0.59
/Pa-ta-ka/	8	Onset	-0.55	0.99	0.85	-0.63	-0.53	0.88	0.78	0.26
	8	Offset	-0.32	0.99	0.99	-0.47	-0.99	0.60	-0.67	0.02

4.2.2 Experiments with HC speakers

Table 4.5 shows the Pearson's correlation between the Bhattacharyya distances and the MDS-UPDRS-III scores considering only HC speakers to train the UBM's using features from voiced/unvoiced segments. It can be observed that the highest correlations per trained models is obtained for the unvoiced segments with respect to the voiced segments considering the read text. For /pa-ta-ka/ the best result was obtained considering features from the voiced segments. In the monologue speech task there is no considerable difference in the results obtained for the voiced/unvoiced segments. It can be observed that the best result was obtained again for the read text ($r = 0.45$). The results for the onset/offset transitions are

Table 4.5 *Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Seg****: Voiced/unvoiced segments.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

UBM trained considering features from voiced/unvoiced segments										
Speech Task	**M**	**Seg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	64	V	0.84	0.19	0.59	-0.22	0.48	-0.48	0.99	0.34
	16	UV	0.45	0.96	-0.82	0.55	0.76	0.60	0.68	0.45
Monologue	256	V	0.74	0.99	0.53	-0.61	-0.50	0.31	-0.84	0.09
	32	UV	0.96	0.43	0.95	-0.41	-0.95	0.33	-0.74	0.08
/Pa-ta-ka/	32	V	0.99	-0.93	0.99	0.55	0.40	-0.94	0.86	0.27
	16	UV	-0.53	-0.65	-0.02	-0.99	0.47	0.92	0.97	0.02

shown in Table 4.6. Again the best result was obtained for the read text considering features from the offset transitions ($r = 0.35$). On the contrary, for the monologue and /pa-ta-ka/ speech task the highest correlations were obtained with features from the onset transitions (monologue: $r = 0.26$, /pa-ta-ka/: $r = 0.28$). Note also that there was an improvement in the results for the monologue for onset/offset transitions with respect to the voiced/unvoiced segments. In this case the best results were obtained for the read text in both experiments. Note also that for this speech task the results obtained in Tables 4.3 and 4.4 are higher than those obtained in Tables 4.5 and 4.6, except for the case of the unvoiced segments.

4.2.3 Experiments with PD patients and HC speakers

For this experiment, the UBM's were trained combining the PD patients and the HC speakers. Table 4.7 shows the Pearson's correlation between the Bhattacharyya distances and the MDS-UPDRS-III scores considering the voiced/unvoiced segments for training. In this case the best results were obtained considering features from the unvoiced segments for the read

Table 4.6 *Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Tran****: Onset/offset transitions.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

UBM trained considering features from onset/offset transitions										
Speech Task	**M**	**Tran**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	256	Onset	0.92	-0.53	0.87	-0.41	0.96	-0.18	0.25	0.27
	32	Offset	-0.53	0.77	0.89	0.08	0.63	0.93	-0.35	0.35
Monologue	64	Onset	-0.10	0.73	0.66	-0.37	0.94	0.93	-0.96	0.26
	16	Offset	0.04	0.90	0.70	-0.55	0.78	0.42	-0.93	0.19
/Pa-ta-ka/	8	Onset	-0.48	0.93	0.93	0.43	0.25	0.45	-0.55	0.28
	16	Offset	-0.80	0.57	0.95	-0.22	0.45	0.99	-0.93	0.14

text and the monologue. For /pa-ta-ka/ the best result was obtained for the voiced segments. Again the best result was obtained for the read text ($r = 0.63$). Note also, that compared to the Tables 4.3 and 4.5 there was a considerable increasing in the performance of the models for the monologue speech task, except when features from the voiced segments are considered for training. Table 4.8 shows the results when the features from the onset/offset

Table 4.7 *Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Seg****: Voiced/unvoiced segments.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

UBM trained considering features from voiced/unvoiced segments										
Speech Task	**M**	**Seg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	64	V	0.96	0.74	-0.47	0.83	0.31	0.71	0.61	0.53
	64	UV	0.87	0.80	0.71	-0.08	0.65	0.73	0.70	0.63
Monologue	512	V	0.76	0.88	0.27	-0.78	-0.86	0.59	-0.90	0.00
	16	UV	0.87	0.95	0.99	0.50	-0.98	0.81	-0.74	0.34
/Pa-ta-ka/	128	V	0.69	-0.79	0.20	0.80	0.99	0.62	0.44	0.42
	4	UV	0.94	-0.99	0.99	-0.13	0.90	0.22	0.36	0.33

transitions are considered for the for training. Again the highest correlation was obtained for the read text considering features from the offset transitions ($r = 0.62$). For the case of the monologue, there was an increase in the correlations for the onset/offset transitions compared with the results obtained with the voiced/unvoiced segments. For /pa-ta-ka/ the correlations obtained with the onset/offset transitions are lower compared with the correlations obtained with voiced/unvoiced segments.

Table 4.8 *Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Tran****: Onset/offset transitions.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between d_{Bha} and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

UBM trained considering features from onset/offset transitions										
Speech Task	**M**	**Tran**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	16	Onset	0.46	0.99	0.88	0.55	0.93	-0.28	-0.55	0.43
	16	Offset	0.77	0.88	0.57	0.51	0.99	0.97	-0.36	0.62
Monologue	32	Onset	0.43	0.52	0.79	0.48	0.99	0.54	-0.69	0.44
	64	Offset	0.41	0.45	0.90	0.53	0.72	0.65	-0.29	0.48
/Pa-ta-ka/	32	Onset	-0.48	0.43	0.95	0.66	-0.25	0.99	-0.89	0.20
	16	Offset	-0.99	0.67	0.99	0.45	-0.17	0.99	-0.98	0.14

In general, the best results were obtained considering features from the unvoiced segments and the offset transitions for the read text speech task for the three UBM's (PD patients, HC speaker, and the combination of PD patients and healthy speakers). In most of the cases the highest correlations were obtained when training the UBM with voiced/unvoiced segments. Conversely, for the monologue the performance of the models improved when the UBM's were trained considering features from the onset/offset transitions, in particular features from the offset transitions as shown in all cases. This can be explained due to the fact that in the monologue each patient said different things and because it is already shown that the motor planning process to produce free-speech is more complex than to read a text [49], thus it seems like this additional difficulty is being captured/modeled with the presented approach. Also, the performance of the models was improved when PD patients and HC speakers were combined to train the models. This can be explained because the GMM-UBM system performs better when it is trained with larger populations. For /pa-ta-ka/ the best results were obtained using the voiced/unvoiced segments, particularly when the UBM is trained using only PD patients and the combination of patients and healthy people. When only the healthy group is considered for training, the performance of the models is similar considering features from voiced/unvoiced segments and onset/offset transitions.

Figures 4.1.A, 4.1.B, and 4.1.C show the best results obtained with the combination of the PD patients and HC speakers used to train the universal model. Figure 4.1.A shows the results considering features from the unvoiced segments for the read text, Figure 4.1.B shows the results considering features from the unvoiced segments for the monologue, and Figure 4.1.C shows the results from the voiced segments for /pa-ta-ka/. The x-axis represents the recording session and the y-axis represents the normalized values of the Bhattacharyya distance and the real MDS-UPDRS-III score. The normalization is performed using the

Figure 4.1.A. Results using unvoiced segments from the read text

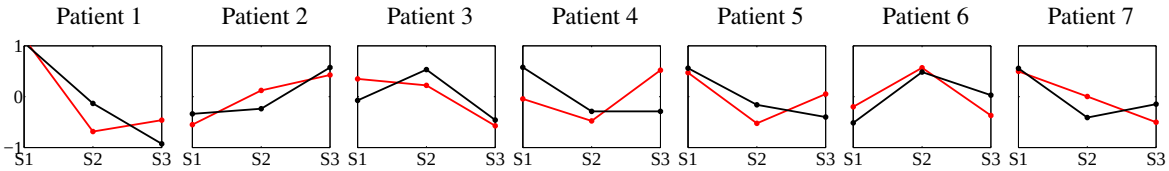

Figure 4.1.B. Results using unvoiced segments from the monologue

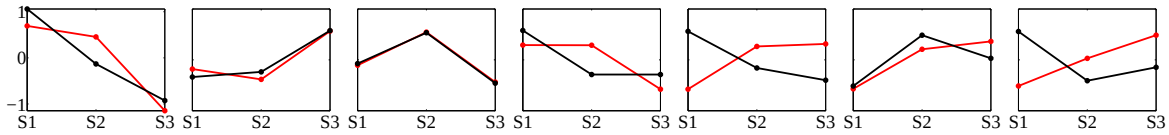

Figure 4.1.C. Results using voiced segments from pataka

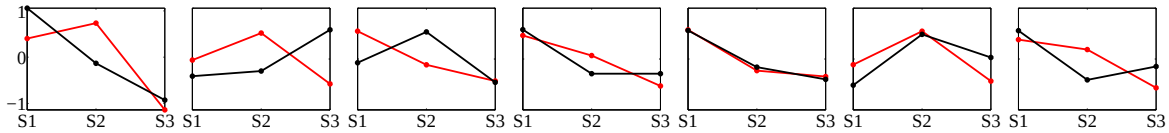

Fig. 4.1 *Normalized scores for each patient considering the Bhattacharyya distance (Red line) and the MDS-UPDRS-III labels (Black line) considering features from the voiced/unvoiced segments when PD patients and HC speaker are combined for training*

z-score approach (MDS-UPDRS-III is represented with the black line and the computed distance with the red line.). This procedure is performed only with the aim of depicting comparable curves (MDS-UPDRS-III and the distances) in the same picture. The distances computed from each user model represent the progression of the disease. Note that the trend of the Bhattacharrya distances follows the trend of the neurological state of the patients. This behavior can be observed clearly for patient 1, patient 2, patient 3 (Figures 4.1.A and 4.1.B), patient 5 (Figure 4.1.C), and patient 6 (Figures 4.1.A and 4.1.C). The results obtained for patient 7 and patient 4 can be explained considering the MDS-UPDRS-III values presented in Table 3.1. For patient 7 there are strong variations in the neurological scores, which indicates that the symptoms of the disease improved considerably from session 1 to session 2. The scores for patient 4 indicate that there are no changes in the neurological state (according to the MDS-UPDRS-III score given by the neurologist). The performance of the speaker models could be improved if information about the pharmacological treatment is considered, i.e, time of the intake of the daily dose, the type of medication, dose of the medications, and others [50].

Figures 4.2.A, 4.2.B, and 4.2.C show the representation of the disease progression considering the features from the onset/offset transitions to train the UBM when PD patients and HC speakers are combined. Figures 4.2.A and 4.2.B. show the results considering features from the offset transitions extracted from the read texts and the monologues, respectively. Figure 4.2.C shows the results obtained using the onset transitions for pataka. For patient 1

Figure 4.2.A. Results using offset transitions from the read text

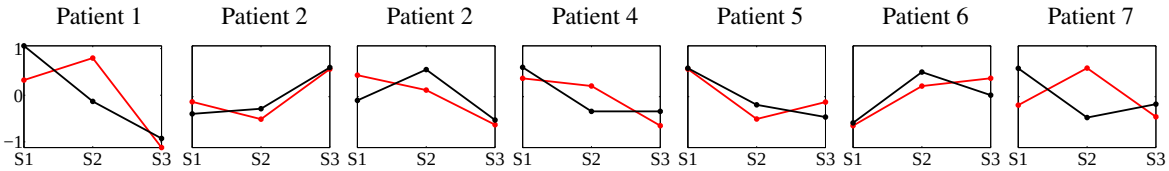

Figure 4.2.B. Results using offset transitions from the monologue

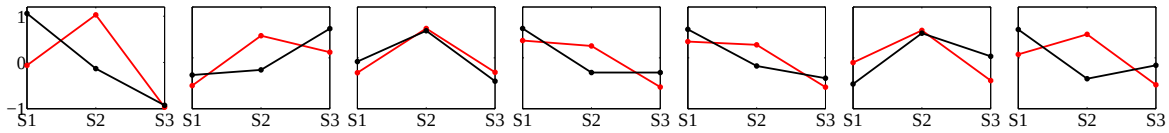

Figure 4.2.C. Results using onset transitions from pataka

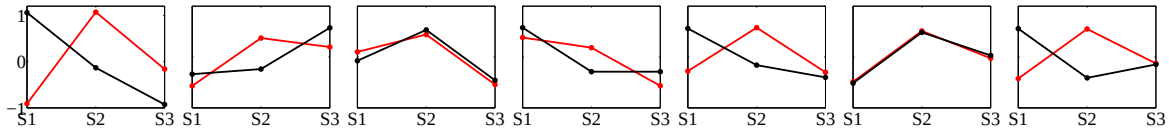

Fig. 4.2 *Normalized scores for each patient considering the Bhattacharyya distance (Red line) and the MDS-UPDRS-III labels (Black line) considering features from the onset/offset transitions when PD patients and HC speaker are combined for training*

the trend of the distances was similar for the three speech tasks. The best result was obtained for the features from the offset transition computed upon the read text ($r = 0.77$). From Tables 4.3, 4.5, 4.7 and Figure 4.1 it can be observed that the best models for patient 1 are obtained considering the information extracted from the voiced/unvoiced segments. Again the for patient 7 and patient 4 the results were not satisfactory. However, note that the trends obtained in Figure 4.1.C and in Figure 4.2 for patient 7 are similar. The same behavior can be observed for patient 4 in Figures 4.1.B, 4.1.C and Figure 4.2. These results seem to indicate that there are changes in speech that are not reflected in the MDS-UPDRS-III scores. However, more data from the patients is needed to validate such a hypothesis.

4.3 Entropy of the models

The entropy of the adapted speaker models is estimated in order to compare the results obtained with the Bhattacharyya distance measure. For these experiments, the entropy of the speaker models obtained with the combination of the PD patients and the healthy group for training are considered. The main reason for this is to test whether it is possible to achieve similar or better results than those obtained with the probabilistic distance. For these experiments, the Rényi entropy (H_{Renyi}, Section 2.7) was estimated for each speaker model. The parameter α of H_{Renyi} was optimize in a grid search process with $\alpha \in \{2,3,4,...,9\}$. The

best results were obtained with $\alpha = 2$. The performance of the models is evaluated computing the Pearson's correlation between the estimated entropies and the MDS-UPDRS-III scores.

Voiced/unvoiced segments:

Table 4.9 shows the obtained results considering the features extracted from the voiced/unvoiced segments. Note that the highest correlation per trained model is obtained for the features extracted from the unvoiced segments. For the case of the read text and the repetition of /pa-ta-ka/ the correlation values are similar to those obtained in Table 4.7. In the case of the monologues, the performance of the models improved from $r = 0.34$ (Table 4.7) up to $r = 0.57$ (Table 4.9). For the voiced segments, the performance of the models is lower which indicates that the features extracted from the unvoiced segments are suitable to track the disease progression from speech. In most of the cases, the results obtained with the Bhattacharyya distance are better (and likely more accurate). This can be likely explained because this distance evaluates the whole structure of the speaker models (means, covariances and weights) and the UBM to estimate the changes in the models.

Table 4.9 *Pearson's correlation between H_{Renyi} ($\alpha = 2$) and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Seg****: Voiced/unvoiced segments.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between H_{Renyi} and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

UBM trained considering features from voiced/unvoiced segments										
Speech Task	**M**	**Seg**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	64	V	-0.72	0.99	-0.98	-0.56	0.06	0.90	0.21	-0.01
	32	UV	0.88	0.72	0.34	0.06	0.98	0.66	0.58	0.60
Monologue	4	V	-0.72	0.74	-0.92	0.06	0.55	0.44	-0.01	0.02
	4	UV	0.75	0.91	0.65	-0.03	0.97	-0.23	0.99	0.57
/Pa-ta-ka/	32	V	-0.90	0.99	0.21	-0.72	0.71	0.52	-0.08	0.10
	4	UV	0.99	-0.46	0.55	0.47	0.99	-0.92	0.47	0.30

Onset/offset transitions:

Table 4.10 shows the results obtained when considering the features extracted from the onset/offset transitions. In this case the best results per speech task were obtained with the features extracted from the offset transitions (read text: $r = 0.41$; monologue: $r = 0.52$; /pa-ta-ka/: $r = 0.30$). Comparing the results presented in Table 4.10 and Table 4.8, the performance of the models is better for the Bhattacharyya distance (read text: $r = 0.62$) than H_{Renyi} (read text: $r = 0.60$). In the case of the monologue, the correlation values are similar using both measures. For the repetition of /pa-ta-ka/ there is an improvement considering features from the offset transitions.

Table 4.10 *Pearson's correlation between H_{Renyi} ($\alpha = 2$) and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Tran****: Onset/offset transitions.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between H_{Renyi} and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model.*

UBM trained considering features from onset/offset transitions										
Speech Task	**M**	**Tran**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	32	Onset	0.87	0.99	-0.89	0.99	-0.97	0.09	0.86	0.28
	256	Offset	0.88	-0.96	0.99	0.98	-0.08	0.25	0.78	0.41
Monologue	2	Onset	0.99	0.15	-0.96	0.06	0.98	0.95	0.79	0.42
	4	Offset	0.97	-0.84	0.34	0.74	0.55	0.99	0.89	0.52
/Pa-ta-ka/	2	Onset	-0.09	0.73	-0.66	0.96	0.20	-0.63	-0.15	0.05
	64	Offset	0.95	-0.95	0.99	0.99	0.97	-0.81	0.14	0.33

From the results obtained in Tables 4.7, 4.8, 4.9, and 4.10 it can be observed that the information from the Bhattacharyya distance can be complemented using a entropy measure in order to improve the models to monitor the disease progression.

4.4 Other measures

Besides the Bhattacharyya distance and the Rényi entropy, the Chernoff distance, the Kullback-Leibler distance, and the Shannon entropy were used to test the performance of the speaker models. See Section 2.6 and Section 2.7 for further details about the computation of these measures. In order to show how the other measures performed with the models, only the best results from Table 4.7 are compared. Table 4.11 shows the results obtained when the UBM is trained considering the features extracted from the unvoiced segments. In the case of the Chernoff distance, the values of β used to test the models range from 10^{-1} to 9×10^{-1} in steps of 10^{-1}, except for $\beta = 0.5$ which is the Bhattacharyya distance (Section 2.6). For the Chernoff distance, the average correlation values obtained with the different values of β are presented. The number of Gaussians used to train the UBM ranges from 2 to 512 in steps of 2^n ($n \in \{1,2,3,...,9\}$). Table 4.11 shows the obtained results. It can be observed that the performance of the models for the Chernoff distance is similar to the results presented in Table 4.7. In the case of the Kullback-Leibler distance, the performance for the read text and the monologue is lower than the Bhattacharyya distance. For the Shannon entropy the performance of the models for /pa-ta-ka/ is better than the results obtained with the Rényi entropy.

Table 4.11 *Pearson's correlation between different measures and the real MDS-UPDRS-III.* ***M****: Number of Gaussians used to train the model.* ***Pi*** *($i \in \{1,...,7\}$): Pearson's correlation between different measures and the real MDS-UPDRS-III score per patient.* ***AVG****: Average value of the correlations per trained model. d_{KL}: Kullback-Leibler distance. d_{Che}: Chernoff distance. H_{Sha}: Shannon entropy.*

UBM trained considering features from unvoiced segments										
Speech Task	**M**	**Measure**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**
Read text	128	d_{KL}	0.15	-0.99	0.99	0.91	0.38	-0.59	0.07	0.13
	64	d_{Cher}	0.77	0.79	0.74	-0.16	0.64	0.78	0.70	0.61
	256	H_{Sha}	0.83	-0.62	0.20	0.50	0.28	-0.05	0.71	0.26
Monologue	2	d_{KL}	-0.79	-0.99	-0.81	0.99	0.80	0.85	0.92	0.14
	2	d_{Cher}	0.65	-0.92	-0.85	0.92	0.57	0.99	0.98	0.33
	128	H_{Sha}	-0.92	0.58	0.13	0.38	-0.96	0.92	0.71	0.12
/Pa-ta-ka/	8	d_{KL}	0.45	-0.62	0.93	0.47	-0.09	0.25	0.68	0.30
	4	d_{Cher}	0.97	-0.95	0.96	-0.09	0.80	0.34	0.37	0.34
	8	H_{Sha}	0.80	0.43	0.80	0.02	0.69	0.94	0.98	0.67

4.5 Analysis of the results

4.5.1 Regression model

The prediction of the neurological state of the patients is performed using a linear ε-SVR. First, the regressor is trained using the features extracted from the voiced/unvoiced segments. From the results, it can be observed that the best performance in the prediction of the disease severity was achieved for the unvoiced segments. Particularly, for the read text and the repetition of /pa-ta-ka/, which indicates that the information provided by the unvoiced segments is more sensitive for dedicate speech tasks than continuous speech (monologue). When features from the onset/offset transitions are considered for training, the performance of the regression models is improved for the three speech tasks. These results indicate that the difficulties of PD patients at the start and stop of the vibration of the vocal folds during speech can be assessed using regression techniques. However, further experiments are required to improved the performance of the models, e.g, develop a more robust regression-based approach.

4.5.2 Individual speaker/patient models

When the UBM is trained using only the information of the PD patients and the features from voiced/unvoiced segments, it can be observed that similar results are achieved comparing

the regressor model to the GMM-UBM-based approach. Again, the best performance is achieved with the dedicated speech tasks. However, the obtained correlations are considerable higher than those obtained with the ε-SVR. Further, the performance using the features from the voiced segments is better for the individual speaker models respect to the regression technique. These results can be explained considering the difference between the SVR and the GMM-UBM. The SVR assess the disease progression considering the correlation between the predicted neurological state of the patients with respect to a previous fitted linear function (Section 2.5). However, the disease progression varies from patient to patient and the trained regression model may not be suitable to detect these variations. Conversely, the GMM-UBM-based approach assess the progression of the disease considering the changes in the speech of each patient. In this case the trained model, i.e, the UBM, is used as the baseline to detect the changes in the speech of the patients considering individual speaker models. Similarly to the SVR, when the features extracted from the onset and offset transitions are considered for training, the performance of the models is improved. Particularly, the Pearson's correlation for the monologue is higher for the onset/offset transitions with respect to the voiced/unvoiced segments. In this case, the results are higher in the monologue considering that this speech task consists of a free dialogue and the problems in the voice of the patients could be more evident when the speech is not planned. Furthermore, it is already documented in the literature that the motor planing to produce free-speech is more complex than to read a text [49]. According to the results obtained in this work, the speech problems on the monologue may be more evident when the speaker is modeled using the GMM-UBM approach. When the UBM is trained using only healthy speakers, the performance of the model is lower compared to the UBM trained only with patients. One hypothesis is that the range of variability in the features for the healthy people is lower compared to the variability in the features extracted from the patients. Since the individual speaker models are tested against the UBM, it is possible that the reduced variability in the feature space produces a reduction in the performance of the models.

The best results are obtained when the UBM is trained with both PD patients and healthy speakers. The highest correlations are achieved with the read text considering features from voiced/unvoiced segments and onset/offset transitions. As it is described in Chapter 3, the read text is a dedicated speech task that is phonetically balanced. Note that the best results were obtained in most cases with the read text (with the SVR approach or with the GMM-UBM approach), which is consistent with the results discussed before. For the monologue, the best results were obtained considering features from the onset/offset transitions when patients are included in the training set. These results indicate that voice problems can be

detected considering onset/offset transitions when the speech is improvised (not planned), mainly because patients have more difficulty articulating words and sentences. This result opens the window to continue working on methods for the automatic and unobtrusive monitoring of PD patients [51]. For the case of /pa-ta-ka/, the best results were also obtained using the voiced/unvoiced features when the patients and healthy speakers are included in the training set. Although the results from the monologue (and the read text) indicate that the patients have problems to start and to stop the movement of the vocal cords, the information extracted from the onset/offset transitions may not be enough to model voice problems with the repetition of /pa-ta-ka/. This can be explained considering that /pa-ta-ka/ is limited to three consonants, which means that a different approach is necessary to model speech problems using this speech task.

Chapter 5

Outlook

The assessment of the progression of Parkinson's disease from speech is addressed in this thesis. The proposed methodology allows to track the neurological state of PD patients over the time. Two different approaches were considered. The first consists of the prediction of the neurological score of the patients. To achieve this, a state-of-the-art regression technique based on a Linear ε-SVR was trained using the speech recordings of the PD patients described in Section 3.1. The trained models are tested using the speech recordings of the 7 patients that participated in the three recording sessions. The patients were evaluated by a neurologist expert according to the MDS-UPDRS-III scale. The performance of the models was evaluated calculating the Pearson's correlation coefficient between the predicted labels and the real MDS-UPDRS-III scores for each of the 7 patients. This experiment is performed in order to validate the features extracted from the voiced/unvoiced segments and the onset/offset transitions. According to the results, it is possible to predict the neurological state of the patients with correlations of up to $r = 0.45$ per speaker when the offset transitions are used to train the model. Additionally, features from the unvoiced segments prove to be more suitable to predict the MDS-UPDRS-III than the features extracted from the voiced segments.

The second approach consists of tracking the disease by using the GMM-UBM-based approach. This method allows to assess disease progression from an specific patient, i.e., modeling his/her disease progression considering individually-adapted models for tracking his/her neurological state over the time. Three different UBMs are trained with features from three different groups of speakers: Parkinson's patients, healthy speakers, and the combination of both. The Bhattacharyya distance between the speaker models and the UBM was computed. One distance value per recording session was calculated per patient. The Pearson's correlation between the Bhattacharrya distances and the MDS-UPDRS-III labels was calculated. The highest correlation values are obtained when PD and HC speakers are

combined in the UBM, indicating that it is worth to include information from control people to improve the results of the predictions. When HC and PD speakers are included in the background model, the distance between the UBM and the model of the adapted speaker highly correlates with the disease progression. This result indicates that the prediction of the neurological state of PD patients can be improved by using information of healthy speakers in the training process. To the best of my knowledge this is the first contribution considering a method to track the neurological state of individual PD patients over the time. This thesis is a step forward in the development of computer aided tools for the continuous and unobtrusive monitoring of people with Parkinson's disease.

Future work should include a more detailed analysis of speech considering the GMM-UBM approach. The voice of PD patients is affected in several dimensions of speech. Phonation, articulation, prosody and intelligibility of the patients are also impaired. Training individual speaker models per speech dimension should provide more information about the voice disorders observed in PD patients.

Additionally, this methodology could be implemented in mobile devices. Continuous monitoring of the patients can be achieved through speech using the technology that already exist and it is at hand of most people by using smartphones, tablets, smartwatch. These devices could be used to record the speech of people unobtrusively, e.g, during a regular phone call. Also, dedicated speech tests can be programmed.

Chapter 6

Summary

In this thesis a new methodology to assess the Parkinson's disease (PD) progression per speaker considering different recording sessions is proposed. The motivation for this work is to develop a methodology for monitoring PD patients. The main reason for this is that neurologist experts rely on medical histories, physical and neurological examinations to assess the patients. However, the motor skills of the patients with PD are impaired, thus to visit a hospital to perform medical screenings and/or assessments is not a straightforward task for them. Additionally, the diagnosis and monitoring of PD symptoms is time-consuming and expensive. For these reasons there is an increasing interest in the research community to develop computer aided systems for monitoring PD patients. There are different approaches proposed in the literature to assess the severity of the disease from speech signals. Most of the reviewed works are focused on the prediction of the neurological state of the patients according to a neurological rating scale. Those models are based on transformations and comparisons performed in groups of speakers; however, none of them has addressed the individual modeling of the speech of each patient to assess the disease progression.

Parkinson's disease affects motor and non-motor activities of patients. The production of speech is a motor activity that involve several muscles and limbs, thus it makes sense to model motor capabilities from speech considering different aspects such as stability in the vocal folds vibration, energy content, articulatory capability, and others. For this reason the articulatory model proposed in [23] is considered here to evaluate speech problems of PD patients. To model the speech problems of the patients, individual speaker models are obtained using the Gaussian Mixture Model-Universal Background Model (GMM-UBM)-based approach. This is a probabilistic modeling method that is commonly used in speaker recognition-based systems. In this work, this approach is used to model the speech of PD patients to assess their disease progression over the time.

Speech signals from sixty four patients and sixty four healthy speakers were collected in three

recording sessions from 2012 to 2015. A subset of seven patients participated in all recording sessions, thus their speech samples were considered to test the speaker models. Three speech tasks were considered for the experiments: the repetition of the word /pa-ta-ka/, a monologue, and the reading of a phonetically balanced text. A professional audio setting was used for the first two recording sessions, and the third session was recorded in non-controlled acoustic conditions using a portable device. For the recordings collected in the third session, the suitability of the captured signals for detecting PD was evaluated considering a Support Vector Classifier-based approach. The obtained results were satisfactory; thus, these speech recordings were considered for the speaker models.

The speaker models are generated from the speech recordings of the seven patients that participated in all sessions. One of the sixty four patients from the database is extracted to be modeled. The individual speaker models are adapted from a background model which is considered as the baseline to assess the disease progression according to its distance to the adapted speaker. The distance between models is computed using a probabilistic dissimilarity measure known as the Bhattacharyya distance. This measure is considered in this thesis, because it evaluates the difference in the weights, means, and covariances of the individual speaker models through the three recording sessions allowing to measure the changes in the speech of the patients.

The speech recordings of the healthy speakers are also considered in the training process to evaluate if there is an improvement on the performance of the models. For each patient three single models are obtained, one per recording session. The disease progression is evaluated by calculating the distance between the universal background model and the speaker model. Finally, the performance of the models is evaluate with the calculation of the Pearson's correlation coefficient between the distance measures estimated for each recording session and the three neurological scores. Considering that the severity of the symptoms get worse over the time, the MDS-UPDRS-III scores are expected to increase through sessions. Furthermore, according to the studies reviewed in the literature the speech of the PD patients is impaired and such an impairment is progressive. If the neurological changes are reflected in the speech of the patients, then the estimated distances should coincide with the trend of the changes in the MDS-UPDRS-III.

The disease progression per patient is evaluated considering two different approaches: a prediction model and the individual speaker models-based approach. The first method consists of a state-of-the-art regression technique based on a linear Support Vector Regressor (SVR). Most of the studies related to monitoring of PD had considered the SVR-based approach to predict the neurological state of the patients according to a standard neurological rating scale. In this thesis, the state of the severity is predicted according to the MDS-UPDRS-III

rating scale. The voiced/unvoiced segments and the onset/offset transitions are extracted and grouped separately. Thus, four linear SVRs are trained to validate the suitability of the articulatory model to evaluate the severity of the disease. From the results obtained with the voiced/unvoiced segments it can be observed that the best performance was obtained with unvoiced segments, particularly for the read text ($r = 0.34$) and the repetition of /pa-ta-ka/ ($r = 0.31$). This results indicate that the information provided by the unvoiced segments is more sensitive for dedicate speech tasks than continuous speech (monologue). When features from the onset/offset transitions are considered for training, the performance of the regression models is improved for the three speech tasks, which partially confirms the hypothesis that the difficulties of PD patients at the start and stop of the vibration of the vocal folds during speech are reflected using the articulatory model.

For the patient/speaker modeling approach three different versions of the UBM were considered: (1) with recordings of a total of sixty three PD patients, (2) with with sixty four healthy speakers, and (3) with both groups of speakers. When the UBM is trained using only the information of the PD patients and the features from voiced/unvoiced segments, similar results are achieved comparing the regressor model to the GMM-UBM-based approach. The best performance is achieved with the read text ($r = 0.45$) and /pa-ta-ka/ ($r = 0.38$). Similarly to the regression-based approach, when the features extracted from the onset and offset transitions are considered for training, the performance of the models is improved. In this case, the results are higher in the monologue ($r = 0.59$) considering that this speech task consists of a free dialogue and the problems in the voice of the patients may be more evident when the speech is not planned.

When the UBM is trained using only healthy speakers, the performance of the model is lower compared to the UBM trained only with patients. One hypothesis is that the range of variability in the features for the healthy people is lower compared to the variability in the features extracted, which affects the computation of the distance between the speaker models and the universal background model.

The best performance of the models are obtained with both PD patients and healthy speakers. The highest correlations are achieved with the read text ($r = 0.62$) considering features from voiced/unvoiced segments and onset/offset transitions. For all of the experiments the performance obtained with the read text is more consistent. This can be explained considering that the read text is a dedicated speech task that is phonetically balanced. For the monologue, the best results were obtained considering features from the onset/offset transitions when patients are included in the training set ($r = 0.59$). These results indicate that voice problems can be detected considering onset/offset transitions when the speech is not planned, mainly because patients have more difficulty articulating words and sentences. For the case of /pa-ta-ka/,

the best results were obtained using the features extracted from the voiced ($r = 0.42$) and unvoiced ($r = 0.33$) segments when the patients and healthy speakers are merged in the training set. Although the results from the monologue (and the read text) indicate that the patients have problems to start and to stop the movement of the vocal cords, the information extracted from the onset/offset transitions may not be enough to model voice problems with the repetition of /pa-ta-ka/. This can be explained considering that /pa-ta-ka/ is limited to three consonants, which means that a different approach is necessary to model speech problems using this speech task.

Regarding the comparison between the regression technique and the proposed approach, i.e, assessment of Parkinson's patients using individual speaker models, it can be observed that the proposed methodology outperforms the state-of-the-art approach. The difference in the results can be explained considering that with the SVR the disease progression is assessed considering the correlation between the predicted neurological state of the patients with respect to a previous fitted linear function that is found considering groups of speakers. However, the disease progression varies from patient to patient. The proposed approach evaluates the progression of the disease considering the changes in the speech of each patient. In this case the trained model is used as the baseline to detect the changes in the speech of the patients considering individual speaker models. The results obtained in this thesis suggest that it is possible to track the disease progression from speech and future work should include a more detailed analysis of speech to validate such a hypothesis.

List of figures

List of tables

References

[1] M. de Rijk, "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts," *Neurology*, vol. 54, no. 5, pp. 21–23, 2000.

[2] M. Trail, C. Fox, L. O. Ramig, S. Sapir, J. Howard, and E. C. Lai, "Speech treatment for Parkinson's disease," *NeuroRehabilitation*, vol. 20, no. 3, pp. 205–221, 2005.

[3] E. Dorsey *et al.*, "Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030," *Neurology*, vol. 68, no. 5, pp. 384–386, 2007.

[4] J. Sánchez, O. Buriticá, D. Pineda, C. Uribe, and L. Palacio, "Prevalence of Parkinson's disease and Parkinsonism in a Colombian population using the capture-recapture method," *International Journal of Neuroscience*, vol. 113, no. 2, pp. 175–182, 2004.

[5] G. Pradilla, B. Vesga, F. León-Sarmiento, and grupo GENECO, "Estudio neuroepidemiológico nacional (EpiNeuro) colombiano," *Revista Panamericana de Salud Pública*, vol. 14, no. 2, pp. 104–111, 2003.

[6] O. Hornykiewicz, "Biochemical aspects of Parkinson's disease," *Neurology*, vol. 51, no. 2 Suppl 2, pp. S2–S9, 1998.

[7] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[8] S. Skodda, W. Visser, and U. Schlegel, "Short- and long-term dopaminergic effects on dysarthria in early Parkinson's disease," *Journal of Neural Transmission*, vol. 117, no. 2, pp. 197–205, 2010.

[9] G. Schultz and M. Grant, "Effects of speech therapy and pharmacologic and surgical treatments on voice and speech in Parkinson's disease: a review of the literature," *Journal of Communication Disorders*, vol. 33, no. 1, pp. 59–88, 2000.

[10] D. G. Theodoros, G. Constantinescu, T. G. Russell, E. C. Ward, S. J. Wilson, and R. Wootton, "Treating the speech disorder in Parkinson's disease online," *Journal of Telemedicine and Telecare*, vol. 12, no. suppl 3, pp. 88–91, 2006.

[11] C. G. Goetz *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[12] A. Tsanas, M. Little, P. E. McSharry, and L. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Transactions on, Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[13] S. Skodda, W. Grönheit, N. Mancinelli, and U. Schlegel, "Progression of voice and speech impairment in the course of Parkinson's disease: a longitudinal study," *Parkinson's Disease*, vol. 2013, 2013, art. ID 389195.

[14] P. Gómez-Vilda, M. C. Vicente-Torcal, J. M. Ferrández-Vicente, A. Álvarez-Marquina, V. Rodellar-Biarge, V. Nieto-Lluis, and R. Martínez-Olalla, *Parkinson's Disease Monitoring from Phonation Biomechanics*. Springer International Publishing, 2015, pp. 238–248.

[15] P. Gómez-Vilda, A. Álvarez-Marquina, V. Rodellar-Biarge, V. Nieto-Lluis, R. Martínez-Olalla, M. C. Vicente-Torcal, and C. Lázaro-Carrascosa, "Monitoring Parkinson's Disease from phonation improvement by Log Likelihood Ratios," in *Proceedings of the Fourth International Work Conference on Bioinspired Intelligence (IWOBI)*, 2015, pp. 105–110.

[16] M. Asgari and I. Shafran, "Extracting Cues from Speech for Predicting Severity of Parkinson's Disease," in *IEEE International Workshop on, MLSP*. IEEE, 2010, pp. 462–467.

[17] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of Parkinson's disease from speech," *Computer Speech and Language*, vol. 29, no. 1, pp. 172–185, 2015.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[19] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition," in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 478–482.

[20] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2014, pp. 342–347.

[21] T. Grósz, R. Busa-Fekete, G. Gosztolya, and L. Tóth, "Assessing the Degree of Nativeness and Parkinson's Condition Using Gaussian Processes and Deep Rectifier Neural Networks," in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 919–923.

[22] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6490–6494.

[23] J. R. Orozco-Arroyave, *Analysis of Speech of People with Parkinson's Disease*. Germany: Logos Verlag Berlin, 2016.

[24] D. Reynolds, *Gaussian Mixture Models*. Boston, MA: Springer US, 2009, pp. 659–663.

[25] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londoño, and E. Nöth, "Automatic Detection of Parkinson's Disease from Continuous Speech Recorded in Non-Controlled Noise Conditions," in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 105–109.

[26] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with parkinson's disease," *Behavioural neurology*, vol. 11, no. 3, pp. 131–137, 1999.

[27] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 246–269, 1969.

[28] D. G. Hanson, B. R. Gerratt, and P. H. Ward, "Cinegraphic observations of laryngeal function in Parkinson's disease," *The Laryngoscope*, vol. 94, no. 3, pp. 348–353, 1984.

[29] H. Ackermann and W. Ziegler, "Articulatory deficits in parkinsonian dysarthria: an acoustic analysis." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 54, no. 12, pp. 1093–1098, 1991.

[30] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in Parkinson's disease," *Journal of Voice*, vol. 25, no. 4, pp. 467–472, 2011.

[31] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[32] K. Stevens, *Acoustic Phonetics*, ser. Current Studies in Linguistics Series. MIT Press, 2000.

[33] J. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *The Journal of the Acoustical Society of America (JASA)*, vol. 139, no. 1, pp. 481–500, 2016.

[34] Petra Zwirner and Thomas Murry and Gayle E. Woodson, "Phonatory function of neurologically impaired patients," *Journal of Communication Disorders*, vol. 24, no. 4, pp. 287–300, 1991.

[35] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 10, pp. 1943–1953, 2006.

[36] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America (JASA)*, vol. 33, no. 2, pp. 248–248, 1961.

[37] T. Jehan, "Creating music by listening," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.

[38] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[39] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[40] V. Vapnik, *The nature of statistical learning theory*. New York: Springer-Verlag, 1995.

[41] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[42] R. O. Duda, P. E. Hart, D. G. Stork *et al.*, *Pattern classification*. Wiley New York, 1973, vol. 2.

[43] P. A. Devijver and J. Kittler, *Pattern recognition theory and applications*. Springer Science & Business Media, 2012, vol. 30.

[44] C. H. You, K. A. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Transactions on, Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1300–1312, 2010.

[45] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[46] A. Rényi *et al.*, "On measures of entropy and information," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1961, pp. 547–561.

[47] J. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/Unvoiced Transitions in Speech as a Potential Bio-Marker to Detect Parkinson's Disease," in *Proceedings Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 95–99.

[48] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[49] D. Van Lancker Sidtis, K. Cameron, and J. J. Sidtis, "Dramatic effects of speech task on motor and linguistic planning in severely dysfluent parkinsonian speech," *Clinical linguistics & phonetics*, vol. 26, no. 8, pp. 695–711, 2012.

[50] B. S. Connolly and A. E. Lang, "Pharmacological treatment of Parkinson disease: a review," *Jama*, vol. 311, no. 16, pp. 1670–1683, 2014.

[51] P. Klumpp, "Implementation of a mobile monitoring application for patients with Parkinson's disease," Master's thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2017.

Appendix A

Publications

The following publications were derived from the development of this work

- **T. Arias-Vergara**, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, and E. Nöth, "Parkinson's disease progression assessment from speech using GMM-UBM", in Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association, 2016, pp. 1933–1937.

- **T. Arias-Vergara**, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Nöth, "Gender-dependent GMM-UBM for tracking Parkinson's disease progression from speech", in Proceedings of the 12th ITG Conference on Speech Communication, Paderborn, Germany, 2016, pp. 259–263.

- J.C. Vasquez-Correa, **T. Arias-Vergara**, J. R. Orozco-Arroyave, J.F. Vargas-Bonilla, and E. Nöth, "Wavelet-Based Time Frequency Representations for Automatic Recognition of Emotions from Speech", in Proceedings of the 12th ITG Conference on Speech Communication, Paderborn, Germany, 2016, pp. 235-239.

- J.C. Vásquez-Correa, **T. Arias-Vergara**, J.R. Orozco-Arroyave, J.F. Vargas-Bonilla, J.D. Arias-Londoño y E. Nöth. "Automatic Detection of Parkinson's Disease from Continuous Speech Recorded in Non-Controlled Noise Conditions". in Proceedings of the 16th International conference from speech and communication association INTERSPEECH, Dresden, Germany, pp.105-109, 2015.

- N. García, **T. Arias-Vergara**, J.R. Orozco-Arroyave, J.F. Vargas-Bonilla."A New Speech Corpus in Spanish for Speaker Verification". Presentado en XXI Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), Bucaramanga, 2016.