*Corresponding author: rafael.orozco@udea.edu.co

Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions

T. Arias-Vergara[1], J.C. Vásquez-Correa[1,2], J. R. Orozco-Arroyave[1,2], and E. Nöth[2]

[1] *GITA research group, Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia*

[2] *Pattern Recognition Lab., Friedrich-Alexander-Universität, Erlangen, Nürnberg*

Abstract

The interest of the research community in the analysis of speech of people suffering from Parkinson's disease has increased in recent years. Most of the studies are focused on developing computer-aided tools for the detection and unobtrusive monitoring the progression of several symptoms of the disease. Different approaches have been proposed to detect several voice impairments in PD patients. Most of the state-of-the-art studies address the task of assessing the neurological state of patients considering information obtained from a set of PD speakers, i.e., most of the reported studies consider regression techniques which are trained with information obtained from a group of patients and assess the neurological state of another group of patients suffering from PD. Such an approach seems to be a good alternative to evaluate the suitability of the models/measures extracted from the speech signals; however, that approach is not appropriate to perform individual monitoring of patients and including information about the progression of the disease in a specific person. Additionally, due to the difficulty of having continue access to PD patients, the number of contributions focused on the automatic monitoring of the patients is reduced. Most of the reported works are based on recordings captured during clinical appointments, i.e., relatively controlled acoustic and recording conditions. In this study we propose a methodology to assess the disease progression considering individual information per patient, i.e., individual speaker models. Two different methods are explored, one is based on the GMM–UBM approach and the other one is based on i–vectors. Both approaches have been successfully applied in speaker identification and verification tasks. In this paper the main hypothesis is that once the speech of a patient is accurately modeled, any change, like those that appear due to the disease progression, will be detected. Speech signals are modeled considering three speech aspects: phonation, articulation, and prosody. The results obtained with the proposed approaches are compared with respect to the traditional framework which is based on regression analysis.

The models are trained considering a set with 100 speakers (50 suffering from PD and 50 healthy speakers). The tests are performed considering two sets with speech recordings captured in real-world acoustic conditions. The first set contains a group of seven speakers recorded several times from 2012 to 2016, i.e., longitudinal recordings. As the acoustic conditions of those recordings were different between sessions, this corpus represents a real-world scenario to study the neurological state of PD patients. The second set is form with recordings of the same group of seven patients recorded in their houses, i.e., at-home recordings, those patients were recorded in 16 sessions during four months, i.e., one day per month, every two hours during eight hours per day. As in the case of the longitudinal recordings, the acoustic conditions were not controlled, thus this set also represents a real-world scenario for the study of the disease progression.

The sets of speech recordings are considered to address two main tasks: (1) the monitoring of the neurological state of each patient according to the motor section of the Movement Disorder Society – Unified Parkinson's Disease Rating Scale (MDS–UPDRS-III) and (2) the monitoring of the dysarthria level of the patients according to a modified version of the Frenchay Dysarthria Assessment (m–FDA). Besides the aforementioned speech recordings, the suitability of the proposed approach is evaluated for tele-monitoring speech disorders developed by PD patients. The system is tested considering different communication channels: Skype®, Hangouts®, mobile phone, and land-line. The results suggest that the i–vector approach is suitable when the acoustic conditions among recording sessions differ. The GMM-UBM approach seems to be more suitable when the acoustic conditions do not change a lot among recording sessions. In general, the results suggest that the proposed approaches are suitable for tele-monitoring the dysarthria level of PD patients.

Keywords: Speech disorders, GMM–UBM, i–vectors, Parkinson's disease, dysarthria, speaker models, longitudinal analysis.

---

1. Introduction

1.1. Motivation

People suffering from PD are characterized by the progressive loss of dopaminergic neurons in the midbrain [1]. PD symptoms include tremor, slow movement, lack of coordination, and speech impairments [2, 3]. Currently, neurologists rely on medical history, physical and neurological examinations to assess the patients. This procedure has two main limitations: (i) it is not objective (the evaluation depends on the doctor's criterion and expertise), and (ii) due to the motor disability of PD patients, to visit a hospital to perform medical screenings and/or assessments is expensive and difficult [4]. Besides such difficulties, the symptoms progress differently among patients, thus it is important to monitor their symptoms individually (per patient) and over long periods of time. Such a monitoring is not feasible if the patient is required to visit the doctor to every screening. The most suitable methods to perform continuous monitoring of the symptoms are based on computer-aided tools. These methods have captured the attention of the research community because they are objective, easy to use, and reproducible. Speech signals are one of the most suitable ways to capture information about the neurological state of PD patients [5, 6, 7]. Studies reported in the state-of-the-art about assessing the neurological state of PD patients from speech signals always consider situations where the acoustic conditions are relatively controlled, i.e., quiet rooms, good/expensive microphones, and direct connection to the recording device. Additionally, the state-of-the-art is mainly based on classical methods to model speech signals, i.e., measurements are extracted from the speech signal and regression methods are used to assess the neurological state of the patient. This paper presents a methodology for the individual monitoring of speech impairments developed by PD patients during the disease progression. The proposed approach overcomes the state-of-the-art in several aspects: (i) the method is based on individual models, which are based on Gaussian Mixture Models - Universal Background Models (GMM–UBM), thus the system performance is adapted to the speech of each patient, (ii) different communication channels are considered including land-lines, mobile phones, Internet-based systems (Skype® and Hangouts®), and traditional recordings performed during a medical appointment. The proposed approach is also tested on two kinds of recordings: (i) signals captured during several recording sessions distributed from 2012 to 2016, and (ii) signals captured in 16 sessions performed in the houses of several patients during 4 months (one day per month, every two hours and during 8 hours). The use of these two recording sets make the experiments reported in this paper highly original and novel, thus we consider that this work is a significant contribution to the development of computer-aided tools to monitor the progression of PD.

1.2. Parkinson's disease: evaluation and monitoring

Neurological evaluation

There is no standard test to diagnose PD. Doctors rely on the clinical history and physical examinations to assess patients. The disease severity is evaluated by neurologist experts by means of several tests. The most widely used are the Unified Parkinson's Disease Rating Scale (UPDRS), the Movement Disorder Society - UPDRS (MDS-UPDRS), and the Hoehn & Yahr (H&Y). The UPDRS was developed to monitor PD-related disability and impairment. It has 42 items to evaluate mentation, behavior, and mood (4 items), activities of daily living (13 items), motor activities (27 items), and motor complications (11 items). Although the scale provides a way to capture multiple aspects of PD, there are several weaknesses such as the absence of screening questions on several important non-motor aspects [8]. The MDS-UPDRS is an updated version of the UPDRS. The scale is divided into four sections: section 1 comprises non-motor experiences (13 items), section 2 includes motor activities of daily living (13 items), section 3 evaluates motor capabilities (33 items), and section 4 considers motor complications (6 items) [9]. The scale has a total of 65 items; however, speech is only considered in one of them. Other scale to assess PD patients is the H&Y, which comprises 5 disease stages ranging from unilateral disease (Stage 1) to being in wheelchair or bedbound (Stage 5). Although its wide acceptance, the scale is biased towards postural instability, thus it has incomplete information about other motor and non-motor impairments [10].

Dysarthria level assessment

There are several scales and clinical methods to evaluate dysarthric speech. One of them is the Frenchay Dysarthria Assessment–2 (FDA–2) [11]. The original version of the FDA–2 considers several factors that are affected in people suffering from dysarthria, such as reflexes, respiration, lips movement, palate movement, laryngeal capability, tongue posture/movement, intelligibility, and others. The FDA–2 requires the patient to visit the examiner, which is not possible in most cases when people suffering from PD are considered. Bering this in mind, it was necessary to develop a modified version of the FDA (m–FDA), which can be administered based on speech signals previously recorded, thus the patient is not required to visit the clinician to be evaluated [12]. The m–FDA considers several aspects of speech: respiration, lips movement, palate/velum movement, larynx, tongue, monotonicity, and intelligibility. Speech impairments are evaluated in a total of 13 items and each of them ranges from 0 (normal or completely healthy) to 4 (very impaired), thus the total score of the scale ranges from 0 to 52.

Assessment of the neurological state from speech

In recent years the research community has been interested in developing methods to assess the neurological state of PD patients from speech. One of the reasons to look for such an aim is to reduce treatment and monitoring costs and another reason is to develop objective tools/systems that help clinicians in the assessment and screening of the patients. In [13] the authors proposed a methodology to assess the UPDRS-III score from speech recordings of 82 subjects. The participants were asked to perform three speech tasks including the sustained phonation of the vowel /a/, the rapid repetition of the syllables (/pa/-/ta/-/ka/), and the reading of three standard texts. The set of features extracted from the speech recordings include pitch, spectral entropy, 13 cepstral coefficients, the number and duration of voiced and unvoiced frames, jitter, shimmer, Harmonic to Noise Ratio (HNR), and the ratio of energy in the first and second harmonics. The set of features was computed separately for each speech task. The UPDRS scores were obtained using two Support Vector Regressor (SVR)-based approaches: (1) ϵ-SVR and (2) ν-SVR. Additionally, different kernels were used to train the SVRs including polynomial, radial basis function, and sigmoid functions. The authors reported that it is possible to estimate the UPDRS-III with a Mean Absolute Error (MAE) of 5.66 using an ε-SVR with a cubic polynomial kernel. Later in [14] the authors compared three regression techniques to assess the UPDRS scores including ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO) regression, and linear SVR. Speech recordings of 168 patients were collected in a single recording session. Besides the features described in [13], the authors added information extracted with the openSMILE toolkit [15]. The authors reported that the neurological state of the patients can be assessed with a MAE of 5.5 considering only PD patients in the training process. Besides, in the INTERSPEECH 2015 Computational Paralinguistic Challenge (ComParE 2015) there was a Parkinson's Condition sub-challenge where the task of neurological state evaluation of PD patients from speech was addressed [16]. Recordings of the 50 patients (25 male, 25 female) included in the PC-GITA database [17] were considered to form the train and development subsets. The test set included a total of 11 patients recorded in non-controlled noise conditions, i.e., not using a sound-proof booth and a professional audio setting. A total of 42 speech tasks were considered. The neurological state of the patients was assessed by a neurologist expert according to the motor section of the MDS-UPDRS (MDS-UPDRS-III). The winners of the challenge reported a Spearman's correlation coefficient of 0.65 between the real MDS-UPDRS-III scores and the estimated values. The authors developed a model based on Deep Rectifier Neural Networks and Gaussian Processes Regression [18]. In [19] the authors presented a methodology to estimate the neurological state of PD patients from speech signals. Recordings of Spanish, German, and Czech PD patients were considered to estimate their neurological state according to the UPDRS-III score. The regression process was performed using a linear ϵ-SVR. Four different speech tasks were considered. The authors applied the articulation model introduced in [20]. The model consists of extracting the energy in the transitions from unvoiced to voiced (onset) and from voiced to unvoiced (offset) segments considering different frequency bands distributed according to the Bark and the Mel scales. Additionally, speech intelligibility was objectively evaluated using the Google Inc.® automatic speech recognition system. According to the authors 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 considering the fusion of articulation and intelligibility measures.

Note that most of the studies in the literature are focused on assessing the neurological state of groups of PD patients. Assessments are performed considering only one recording session, thus the disease progression is not evaluated/modeled. The next subsection presents the most recent contributions of the research community to perform longitudinal evaluations, i.e., longitudinal monitoring, of patients suffering from PD considering several recording sessions.

Longitudinal monitoring of PD from speech

There are several studies about automatic monitoring of PD symptoms from speech considering different recording sessions distributed over a period of time. In [5] the authors considered recordings of sustained vowels to estimate the disease progression. The signals were modeled using several acoustic measures including jitter, shimmer, Noise to Harmonic Ratio (NHR), HNR, Relative Amplitude Perturbation, Period Perturbation Quotient, Amplitude Perturbation Quotient, Recurrence Period Density Entropy, Detented Fluctuation Analysis, and Pitch Period Entropy. The UPDRS-III scores were assessed using three linear regression techniques: Least Squares (LS), Iteratively Re-weighted Least Squares, and LASSO. The Classification And Regression Trees (CARTs) approach was also applied. 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 along the study, thus the weekly UPDRS scores were obtained by the authors using a piecewise linear interpolation. The performance of the regression techniques was evaluated using the MAE. The authors reported that the CARTs is the best approach with a MAE of 7.5 points in the evaluation of the total value of the UPDRS scale. The scores of the motor section in the UPDRS (UPDRS-III) were estimated with a MAE of 6 points. This study was one of the first reporting results of PD severity assessment from speech. However, the authors were not aware of the speaker independence because their experiments mixed recordings of the test and train sets, thus the reported results are highly optimistic and biased. The progression of speech impairments in a longitudinal study is presented in [6]. The speech of 80 PD patients (48 male, 32 female) was recorded from 2002 to 2012 in two 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 in the perceptual evaluation: voice, articulation, prosody, and fluency. These aspects were used by the authors to describe motor speech disorders suffered by PD patients. Additionally, an acoustic analysis was performed to describe these speech aspects. Voice was modeled with a set of features including jitter, shimmer, NHR, and average of the 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 compared the extracted features in the first and the second session. The authors found significant differences for shimmer, NHR, NSR, pause ratio, and VAI when features extracted from the first session are compared with respect to the same features extracted from the second session. According to the authors, the results are not conclusive due to some methodological limitations like the long period of time between the two recording sessions. A study for the monitoring of PD progression is also presented in [21]. 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. Sustained phonations of the vowel /a/ were 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 the features: (i) vocal tract inversion using a lattice adaptive filter and (ii) biomechanical inversion of a 2-mass model of the vocal folds. Features from the 50 male healthy controls (HC) were used as baseline to describe the normal state of the speech. During the recording sessions the patients continued their pharmacological treatment and received speech therapy. Each patient was evaluated according to the 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. The authors suggest to define different time intervals between evaluations to obtain more conclusive results. Similarly, in [22], the authors proposed the Log Likelihood Improvement Ratio (LLIR) as a metric to compare speech recordings of eight male PD patients captured in four recording sessions. The patients followed pharmacological treatment and received speech therapy. The aim of the study was to detect changes in the voice before and after the treatment using the same features described in [21]. The authors reported that LLIR is a good metric to detect changes in phonation when the patient is under treatment. Although the authors detected changes in phonation measures, it is not clear whether the same approach is suitable to detect changes in the general neurological state of PD patients. One of the main constraints of addressing longitudinal studies with PD patients is to have continuous contact with them. Thanks to the strong relation of our Lab with the Parkinson's Foundation in Medellín (goo.gl/ihwjLy) we have had continuous contact with Parkinson's patients and they have been actively collaborating in our research activities. In [23] preliminary results using the GMM-UBM approach to model speech impairments developed by seven PD patients were presented. The speech of these patients was captured in several recording sessions between 2012 and 2015. The results of that study motivated us to continue addressing research in individual speaker model methods to monitor symptoms of PD patients. Recently, in [24] the i-vector approach was applied to assess the neurological state of a group with 50 PD patients. Similarly, in [25] speech impairments of PD patients speaking three different languages (Spanish, German, and Czech) were evaluated considering the i-vector approach. The results indicate that this method is suitable to be applied in different languages.

Although the results were promising, those studies were focused on evaluating correlations between a given clinical scale (MDS-UPDRS-III or m-FDA) and the result of a model. In this paper we decided to continue working on this topic but applying the GMM-UBM and i-vector approaches for the individual monitoring of the progression of speech impairments developed by PD patients.

Parkinson's speech evaluation considering non-controlled acoustic conditions

The analysis of PD from voice signals recorded in different acoustic conditions has not been extensively addressed in the literature. In [26], speech recordings of 52 PD patients are transmitted over a simulated mobile telephone network. The authors aimed to estimate the UPDRS scores considering features extracted from sustained phonations of the vowel /a/. Although the aim was very interesting and revolutionary by that time, the results reported in the study were biased because the authors mixed recordings of train and test speakers into the same set, thus the main question regarding the suitability of voice analysis for PD detection remained unanswered. Additionally, besides the necessity of assuring the speaker independence, experiments with continuous speech signals are required in order to extend the application of those approaches to real-world scenarios. Recently, in [27], the effects of background noise, different distortion levels, and telephone codecs were evaluated in the automatic classification of PD vs. HC speakers. The authors concluded that background noise has the strongest effect in the classification accuracy. The effect of telephone channels was not critical, except for the mobile channel, where the low bit-rate codecs caused an important reduction in the classification accuracy.

Contribution of this study

This paper considers speech signals of people suffering from PD recorded during several sessions from 2012 to 2016, i.e., longitudinal study. As a group of speakers is recorded several times, those recordings are suitable to develop a system to model individual changes in the speech of PD patients. Acoustic conditions of those recordings were different between sessions, thus this corpus represents a real-world scenario to study the neurological state of PD patients from speech in real acoustic conditions. Two approaches are explored here, one is based on GMM–UBMs and the other one is based on i–vectors. Both methods are trained considering different aspects of speech: phonation, articulation, and prosody. Additionally, in order to assess the suitability of the approaches in different acoustic and communication conditions, five different communication channels are considered: sound proof booth, Skype®, Google Hangouts®, land-line, and mobile phone. Besides those channels, the proposed approach is tested upon recordings captured in the house of the patients (the same group that is considered in the longitudinal experiments). Those patients were recorded in 16 sessions during four months, i.e., one day per month, every two hours during eight hours per day. As in the case of the longitudinal recordings, the acoustic conditions were not controlled, thus this set represents a real-world scenario for the study of the neurological state of PD patients. To the best of our knowledge this is the first study introducing and testing individual speaker models to monitor PD progression considering speech signals captured with different communication channels/codecs, and at-home recordings.

2. Materials and methods

2.1. Datasets

Three datasets are considered in this study, one is used to train the models and the other two sets are considered to test.

Training set: This is formed with a subset of the PC-GITA database [17] which originally consists of 100 speakers (50 PD patients and 50 HC). The subset includes all of the 50 healthy speakers and 44 PD patients. The remaining 6 speakers are included in the test sets because they participated in further recording sessions and we did not want to lose the chance of including them in individual speaker models. None of the participants in the HC group has history of symptoms related to PD or any other kind of movement and mental disorder. All of the speakers in PC-GITA were recorded in a sound-proof booth with a sampling frequency of 44.1 kHz with a resolution of 16 bits. Different acoustic conditions are tested. The original signals were transmitted and re-captured using four communication systems: Skype®, Google Hangouts®, a landline, and a mobile phone.

All of the PD patients in the training set were evaluated by a neurologist expert according to the MDS-UPDRS-III (due to cost-related reasons healthy speakers were not considered for neurological evaluations). Additionally, the dysarthria level of the patients and the healthy speakers was evaluated by expert phoniatricians according to the m-FDA [28]. The labeling process of the speech recordings was performed by three phoniatricians who were asked to agree on the evaluation of the first ten speakers at the beginning of the process. The remaining recordings were independently evaluated per each phoniatrician. The inter–rater reliability is 0.86. The statistical difference among labels per class (PD and HC) is evaluated by means of the F-statistics of an analysis of variance (ANOVA) test and the results show significant differences between the m–FDA labels of PD and HC speakers, i.e., F = 175.49, $p < 0.001$ for all speakers, F = 66.81, $p < 0.001$ for female speakers, and F = 52.13, $p < 0.001$ for male speakers. Table 1 summarizes the information of speakers in the training set.

Table 1: Description of the training set. **PD patients**: Parkinson's disease patients. **HC**: healthy controls.

	PD patients		**Healthy speakers**	
	male	**female**	**male**	**female**
Number of speakers	22	22	25	25
Age [years] (mean ± standard deviation)	61.3 ± 12.3	61.9 ± 7.3	60.5 ± 11.4	61.4 ± 6.9
Range of age [years]	33–81	49–75	31–86	49–76
Disease duration [years] (mean ± standard deviation)	9.2 ± 6.0	13.0 ± 12.0		
Range of disease duration [years]	0.4–20	1–43		
m-FDA (mean ± standard deviation)	31.2 ± 8.1	32.0 ± 10.1	7.6 ± 7.3	5.1 ± 9.1
Range of m-FDA	17–41	13–51	0–29	0–25
MDS-UPDRS-III (mean ± standard deviation)	40.7 ± 21.5	37.5 ± 15.2		
Range of the MDS-UPDRS-III scores	9–92	19–71		
Average duration of the monologues (in seconds)	47.2 ± 26.4	41.5 ± 20.6	43.1 ± 30.9	54.4 ± 27.3
Average duration of the read texts (in seconds)	18.6 ± 5.9	18.6 ± 6.9	17.5 ± 3.2	18.3 ± 4.2

Longitudinal test set: Speech recordings of 7 patients were collected in five recording sessions from 2012 to 2016. In 2012 (June), 2014 (June), 2015 (February), 2015 (August), and 2016 (February). A professional audio setting was used for the first two sessions. The first recording session includes those six patients who were excluded from PC-GITA to form the training set. An additional speaker who was not part of PC-GITA is also included in this longitudinal set. The speakers in this longitudinal set were recorded in non-controlled acoustic conditions using the device presented in [29]. Six of the seven patients participated in all the sessions. The average duration of the monologues and the read texts were 110.2 ± 42.9 seconds and 17.2 ± 3.8 seconds, respectively. The MDS-UPDRS-III labels of the third recording session (LS3) are not available. Table 2 indicates the MDS-UPDRS-III and the m-FDA labels assigned to the patients of the longitudinal test set. Age and gender are also provided.

Table 2: General information of patients in the longitudinal test set. **LSi:** i-th longitudinal session (**LSi**,i $\in \{1, 2, ..., 5\}$).

			MDS-UPDRS-III					**m-FDA (longitudinal)**				
Patients (Pi)	**Age**	**Gender**	**LS1**	**LS2**	**LS3**	**LS4**	**LS5**	**LS1**	**LS2**	**LS3**	**LS4**	**LS5**
P1	70	M	14	25	–	7	15	37	22	18	23	31
P2	57	M	–	58	–	63	51	–	34	25	34	35
P3	67	M	28	19	–	13	24	31	15	17	16	23
P4	59	F	41	35	–	33	33	29	39	24	21	40
P5	56	F	29	26	–	26	30	23	26	16	16	14
P6	52	F	38	49	–	44	45	14	20	1	12	15
P7	61	M	6	8	–	24	21	21	36	12	13	17

At-home test set: The same group of seven patients considered in the longitudinal test set was recorded four times per day (every two hours), once per month during four months. Thus, there is a total of 16 recording sessions per patient. The participants were recorded in their homes with the same device used for the longitudinal test set [29]. The average duration of the monologues and the read texts were 119.2 ± 57.2 seconds and 18.2 ± 4.1 seconds, respectively. As it was not possible to have a neurologist expert during all day long with each patient, the at-home test set does not have MDS-UPDRS-III scores. The speech recordings of this set were evaluated by one of the phoniatricians who participated in the labeling process with the m-FDA scale. Table 3 indicates the dysarthria scores of the patients in the at-home test set.

Table 3: Dysarthria scores of the at-home test set. **Hi**, i $\in \{1, 2, .., 16\}$: m–FDA scores of the sixteen recording sessions.

			m-FDA (At-home)															
Patients (Pi)	**Age**	**Gender**	**H1**	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**	**H12**	**H13**	**H14**	**H15**	**H16**
P1	70	M	25	25	25	23	21	27	27	27	27	27	27	27	22	21	22	22
P2	57	M	37	38	35	35	35	38	35	37	37	27	39	37	36	36	37	39
P3	67	M	23	23	23	6	23	14	12	12	12	17	22	23	28	22	16	16
P4	59	F	33	34	34	34	33	33	33	33	33	34	36	36	41	41	41	41
P5	56	F	27	25	25	25	31	29	29	29	29	29	31	31	39	39	37	39
P6	52	F	13	13	13	13	13	13	13	13	15	15	15	15	16	14	14	14
P7	61	M	23	24	24	23	26	26	25	25	26	26	26	26	26	25	24	24

2.2. Methods

Figure 1 summarizes the stages of the proposed methodology. In the first stage, one patient is selected to be modeled/tested and the remaining speakers are considered for training the reference model. Afterwards, voiced/unvoiced segments and onset/offset transitions are segmented from the speech recordings. Different features are computed upon the segments depending on the modeled speech aspect (phonation, articulation, or prosody). The measures extracted from the training set are used to create the UBM. The set of features extracted from the recordings of the patient who is being monitored is used to obtain an individual model which is adapted from the UBM. Finally, the disease progression (in terms of the neurological state or the dysarthria level) is evaluated calculating the distance between the UBM and the speaker model. The proposed approach is compared with respect to a regression model, which has been the typical way of addressing the problem introduced in this paper. The next subsections provide details of each stage of the methodology.

2.2.1. Segmentation

The speech production mechanism involves different subsystems mainly formed with muscles and limbs in the vocal tract. The phonatory subsystem is in charge of producing voiced sounds by taking the airflow from the lungs to make the vocal fold vibrate. The articulation subsystem involves the movement and control of different limbs and muscles including tongue, jaw, lips, and velum. This subsystem is involved in the production of vowels and unvoiced sounds like plosive and nasal consonants. When unvoiced sounds are produced there is no vocal fold vibration and those sounds are generated by turbulent

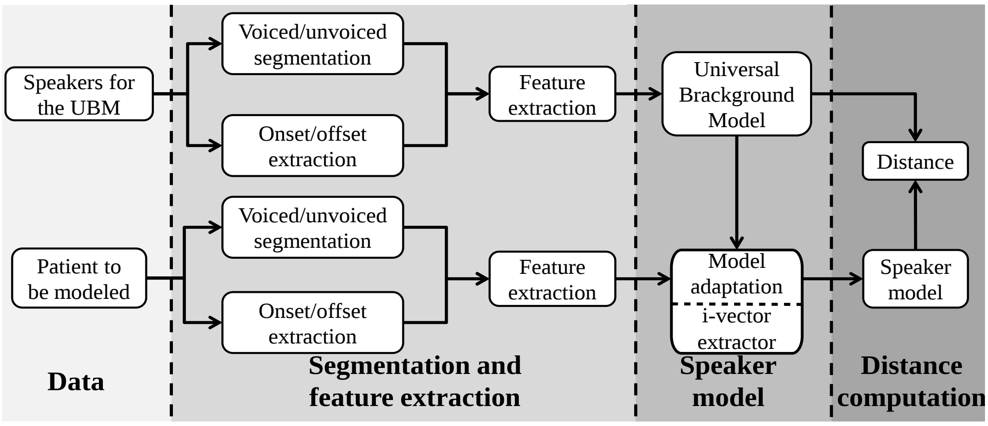


Figure 1: *General methodology to build the speaker models and estimate their degree of impairment.*

airflow at a constriction in the vocal tract. During the production of the voiced segments the vibration of the vocal fold follows four stages in one cycle: (1) closed, (2) opening, (3) open, and (4) closing. Figure 2 shows these stages.

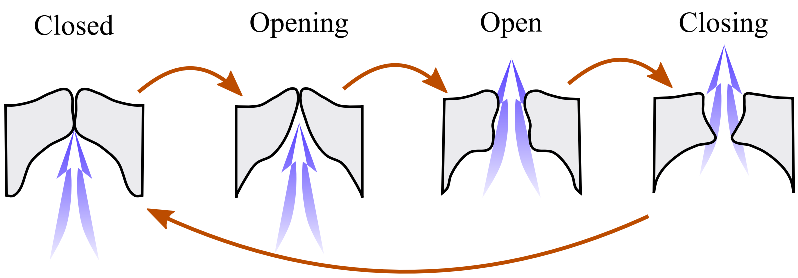


Figure 2: *Vocal folds vibration pattern during voiced segments (Based on a figure found in [30]).*

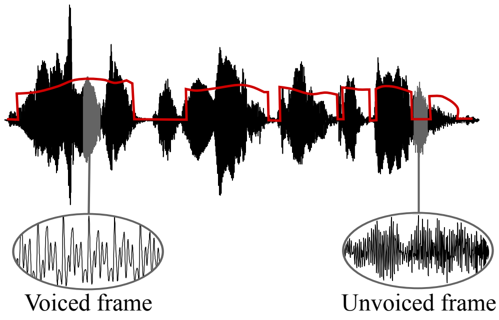
There are several frequency and amplitude perturbation patterns which are observable during the production of vocal sounds. Those perturbations result from different factors such as the vocal fold asymmetry, involuntary movements at the larynx (neurogenic factors), and fluctuations of the airflow and subglottal pressure [30]. On the other hand, the unvoiced segments are produced by a total constriction at certain place in the vocal tract resulting in the interruption of the airflow. Unvoiced sounds are also produced by narrowing the air path producing turbulent airflow which creates noise-like signals [31].

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 [7]. Figure 3.A shows the pitch contour (red line) obtained from a voice recording. It can be observed that voiced segments are quasi-periodic signals, while the unvoiced segments are noise-like signals.

Onset and offset transitions are considered to model difficulties of the PD patients to start and to stop a movement like the vocal fold vibration (Figure 3.B) [20]. Those transitions are produced by the combination of different sounds during the continuous speech production.

Figure 3.A. Voiced/unvoiced segments

Figure 3.B. Onset/offset transitions



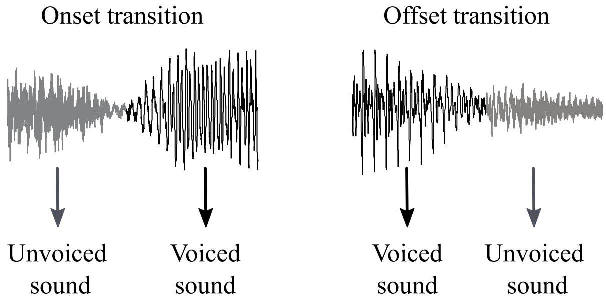


Figure 3: *(A) Pitch contour (red line) and voiced/unvoiced short time windows extracted from a speech signal. (B) Onset and offset transition frames.*

2.2.2. Feature extraction

Voiced/unvoiced segments and onset/offset transitions are used to analyze speech impairments in PD patients considering phonation, prosody, and articulation measures. Features extracted from the voiced segments are considered to model the temporal and amplitude variation of the vocal fold vibration. Prosodic impairments are modeled considering pitch and energy contours extracted from the voiced segments. Articulation impairments are modeled considering spectral measures and the energy content of the onset/offset transitions. Phonation and articulation features were extracted using the software presented in [28].

Phonation features

The evaluation of phonation impairments in continuous speech is performed extracting voiced segments from the monologues and the read texts. The set of features include temporal and amplitude variations of the pitch period, i.e., jitter and shimmer, respectively. Further, the first and second derivatives of the pitch contour are considered to analyze the temporal variability of the fundamental frequency.

Prosodic features

Prosody is analyzed considering pitch and energy-based features extracted from the voiced segments. The set of features is computed based on the methodology presented in [32]. A 5-degree polynomial function is fit to the pitch and energy contours, separately. Then, the 6 coefficients of each fitted curve are used to model prosodic features such as the mean pitch/energy of the voiced segment, the slope of the contour, and the curvature of the pitch/energy contours. Additionally, the duration of each voiced segment is considered to form a 13-dimensional feature vector.

Articulation features

The articulatory capability of the patients is evaluated considering information from the onset/offset transitions. The set of features includes 12 Mel-Frequency Cepstral Coefficients (MFCCs), which comprises a smoothed representation of the speech spectrum considering information of the human auditory system, mainly the critical-band frequency resolution.

These features are widely used to model articulatory problems in the vocal tract [33]. Additionally, in order to incorporate valuable information evidenced in psychoacoustic experiments [34, 35, 36], the log energy of the signal distributed in 22 Bark bands are extracted from the onset/offset transitions.

2.2.3. Regression model

The baseline to estimate the disease severity according to the m–FDA and MDS-UPDRS-III scales (y) is calculated based on a radial basis Support Vector Regressor (SVR) with an ε-insensitive loss function, i.e., ε–SVR. The estimation ($\widehat{y}$) is measured with an ε-insensitive loss function $L(y, \widehat{y})$, which ensures the existence of the global minimum, and it is computed with Equation 1.

$$L(y, \widehat{y})) = \begin{cases} 0 & \text{if } |y - \widehat{y}| \leq \varepsilon \\ |y - \widehat{y}| - \varepsilon & \text{otherwise} \end{cases} \tag{1}$$

The feature vectors $\boldsymbol{x}$ are mapped into a m-dimensional feature space using a linear kernel $g(\boldsymbol{x})$. The estimated values $\widehat{y}$, with weights ω, and bias b, are estimated using Equation 2.

$$\widehat{y} = \sum_{j=1}^{m} \omega_j g_j(\boldsymbol{x}) + b \tag{2}$$

The performance is evaluated using the Spearman's correlation coefficient between the estimated values and the clinical labels.

2.2.4. Speaker models

This paper introduces the use of Gaussian Mixture Models – Universal Background Models (GMM–UBM) to quantify the disease progression. These kind of models have been successfully used in speaker recognition and verification tasks. The main hypothesis in this work is that if the speech of a PD patient is changing due to the disease progression, such a change should be modeled and quantified by a GMM–UBM system. In this case, instead of comparing the speech of one speaker with respect to a different one or to a group of speakers, the idea is to compare the speech of one patient recorded in one moment with respect to the speech of the same patient recorded in a different moment. As PD is progressive and affects speech, any change in the speech production should be captured by the proposed model.

The GMM–based systems are capable of representing arbitrary probabilistic densities. GMMs are parametric probabilistic models represented as a weighted sum of M Gaussian densities. For a D-dimensional feature vector $\boldsymbol{x}$ a GMM is defined as:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} \omega_i p_i(\boldsymbol{x}) \tag{3}$$

The Gaussian densities $p_i(\boldsymbol{x})$ are parameterized by the mixture weights ω_i, a D×1 mean vector $\boldsymbol{\mu_i}$, and a D×D covariance matrix $\boldsymbol{\Sigma_i}$ [37]. The parameters of the density models can be denoted as $\lambda = (\omega_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ and the Gaussian densities as

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma_i}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i^{-1}} (\boldsymbol{x} - \boldsymbol{\mu_i})\} \tag{4}$$

In speech processing GMMs are used to represent the distribution of feature vectors extracted 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 $p(\boldsymbol{X}_{UBM}|\lambda)$ is known as the maximum likelihood function that better represents the population of speakers, where $\boldsymbol{X}_{UBM}$ 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 $k+1$, $p(\boldsymbol{X}|\lambda^{(k+1)}) > p(\boldsymbol{X}|\lambda^{(k)})$. The model of the test speaker is derived from the population of speakers by adapting the parameters of the UBM using the Maximum A Posteriori (MAP) adaptation.

2.3. Identity vectors

This is another way of creating speaker models. This approach has been extensively used in speaker verification and identification tasks. An i–vector is defined in a single space called total variability space which contains both the speaker and channel variabilities simultaneously [38]. The use of a total variability matrix was motivated by [39] after it was showed that channel factors in Joint Factor Analysis (JFA) also contain information about speakers.

In this approach the speaker supervector $\boldsymbol{M}$ is given by:

$$\boldsymbol{M} = \mathbf{m} + \boldsymbol{T}\boldsymbol{\omega} \tag{5}$$

where $\mathbf{m}$ is the channel- and speaker-independent super-vector (usually the super-vector of the UBM), $\boldsymbol{T}$ is the total variability matrix which is trained in the same way as the eigen-voice $\boldsymbol{V}$ matrix, and the components of $\boldsymbol{\omega}$ are the total factors, and $\boldsymbol{\omega}$ itself is known as the identity vector or i–vector.

According to [38], $\boldsymbol{\omega}$ is defined by its posterior distribution conditioned to the Baum-Welch statistics. Given a sequence of L frames $\{y_1, y_2, \ldots, y_L\}$ and a UBM Ω composed of C mixture components, the Baum-Welch statistics N_c and F_c of utterance u are given by:

$$N_c = \sum_{t=1}^{L} P(c|y_t, \Omega) \tag{6}$$

$$F_c = \sum_{t=1}^{L} P(c|y_t, \Omega) y_t \tag{7}$$

where $c = 1, \ldots, C$ is the Gaussian index and $P(c|\boldsymbol{y_t}, \Omega)$ is the posterior probability of mixture component c generating the vector $\boldsymbol{y_t}$.

The first-order Baum-Welch statistic centralized around the mean of the UBM mixture component c (i.e., m_c) is given by:

$$\tilde{F}_c = \sum_{t=1}^{L} P(c|y_t, \Omega)(y_t - m_c) \tag{8}$$

Then, the identity vector $\boldsymbol{\omega}$ for a given utterance u can be found as follows:

$$\boldsymbol{\omega} = (I + T^t\Sigma^{-1}N(u)T)^{-1}T^t\Sigma^{-1}\tilde{F}(u) \tag{9}$$

where $N(u)$ is a diagonal matrix whose diagonal blocks are $N_c I$, $\hat{F}(u)$ is a supervector that concatenates all of the first-order Baum-Welch statistics $\tilde{F}_c$ for a given utterance u, and Σ models the residual variability not captured by the total variability matrix $\boldsymbol{T}$.

Distance computation: GMM–UBM

The neurological state and the dysarthria level 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 features described in Section 2.2.2. One way to assess the changes in the speech of the patients consists of calculating the Bhattacharyya distance. It is a probabilistic measure that considers the weights, the mean vectors, and the covariance matrices of the UBM and the speaker models. When GMM models are considered, the Bhatthacharyya distance can be estimated as:

$$d_{Bha} = \frac{1}{8}\sum_{i=1}^{M}\left\{(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T\left[\frac{\widehat{\boldsymbol{\Sigma}_i} + \boldsymbol{\Sigma}_i}{2}\right]^{-1}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\right\} + \frac{1}{2}\sum_{i=1}^{M}\left[\ln\frac{\left|\frac{\widehat{\boldsymbol{\Sigma}_i}+\boldsymbol{\Sigma}_i}{2}\right|}{\sqrt{\left|\widehat{\boldsymbol{\Sigma}_i}\right|\left|\boldsymbol{\Sigma}_i\right|}}\right] - \omega_{\text{Bha}} \tag{10}$$

Here $\omega_{Bha} = \frac{1}{2}\sum_{i=1}^{M}\ln(\widehat{\omega_i}\omega_i)$ is the mixture weight measure, $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}_i}$ are the mean vector and the covariance matrix of the UBM, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix of the speaker model [40]. The disease progression is evaluated by calculating the Bhattacharyya distance between the UBM and the speaker model. The details of the procedure are depicted in Figure 4.

Distance computation: i–vectors

Similar to the GMM–UBM approach, i–vectors are used to assess the dysarthria level and neurological state of the patients over the time. In this case the measure to estimate the disease progression is the dot product (Equation 11) between the i–vectors extracted from patients and speakers from the UBM.

$$d_{cos} = \frac{\langle \omega_{UBM}, \omega_{SPK} \rangle}{\|\omega_{UBM}\|\|\omega_{SPK}\|} \tag{11}$$

where ω_{UBM} and ω_{SPK} are the i–vectors extracted from the UBM and each patient, respectively. ω_{UBM} is the average i–vector calculated considering the i–vectors of all of the speakers in the UBM. The details of the procedure are depicted in Figure 5.

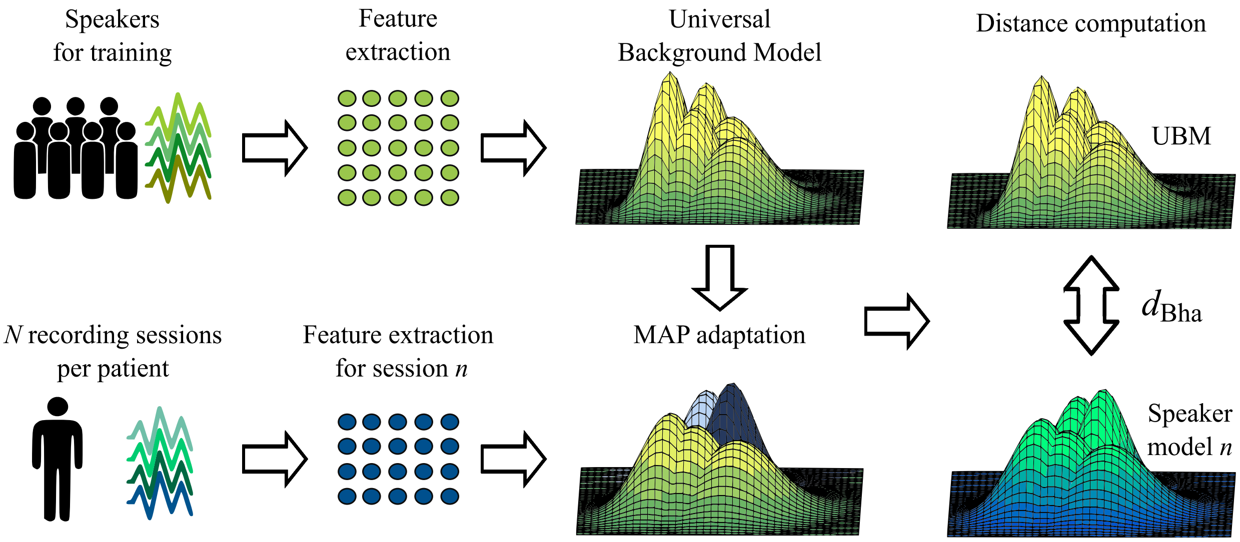


Figure 4: *Speaker modeling. PD progression in N recording sessions per patient: $n \in \{1, 2, 3, ..., N\}$.*

Distances transformed to similarity measures

The speaker models are created with the aim of quantifying changes of two clinical variables over the time: (i) the neurological state according to the MDS-UPDR-III scale, and (ii) the dysarthria level according to the m–FDA score. The performance of the proposed models is evaluated with the Spearman's and Pearson's correlation coefficients calculated between the estimated distance (Bhattacharyya or dot product) and the corresponding scores (MDS-UPDRS-III or m–FDA). Those correlation coefficients measure the relationship between two variables in the interval $[-1, 1]$, were the extreme values represent maximum correlation. The computed distances per speaker model are transformed into similarity measures using Equation 12 [41].

$$s_i = 1 - d_i \tag{12}$$

where $d_i, i \in \{1, 2, 3, .., 7\}$ are the distances computed per speaker model, using the GMM–UBM and i–vectors approaches. This transformation is performed to obtain positive values in all of the cases.

The three speech aspects introduced in Section 2.2.2 (phonation, articulation, and prosody) are considered per patient, thus for each speaker three different distances are computed. Those distances are integrated in the multi-aspect coefficient ξ which is proposed in this paper as indicated in Equation 13

$$\xi_i = \frac{1}{1 + \alpha \mathbf{phon_i} + \beta \mathbf{pro_i} + \theta \mathbf{art_i}} \tag{13}$$

where $\mathbf{phon_i}$, $\mathbf{pro_i}$, and $\mathbf{art_i}$ are the distances corresponding to the phonation, prosody, and articulation aspects, respectively for the patient $\mathbf{i}$. α, β, and θ are the weights of each aspect and are computed as follows: the distances of six of the seven test speakers are considered to train a linear regressor. The parameter associated to the regression line is

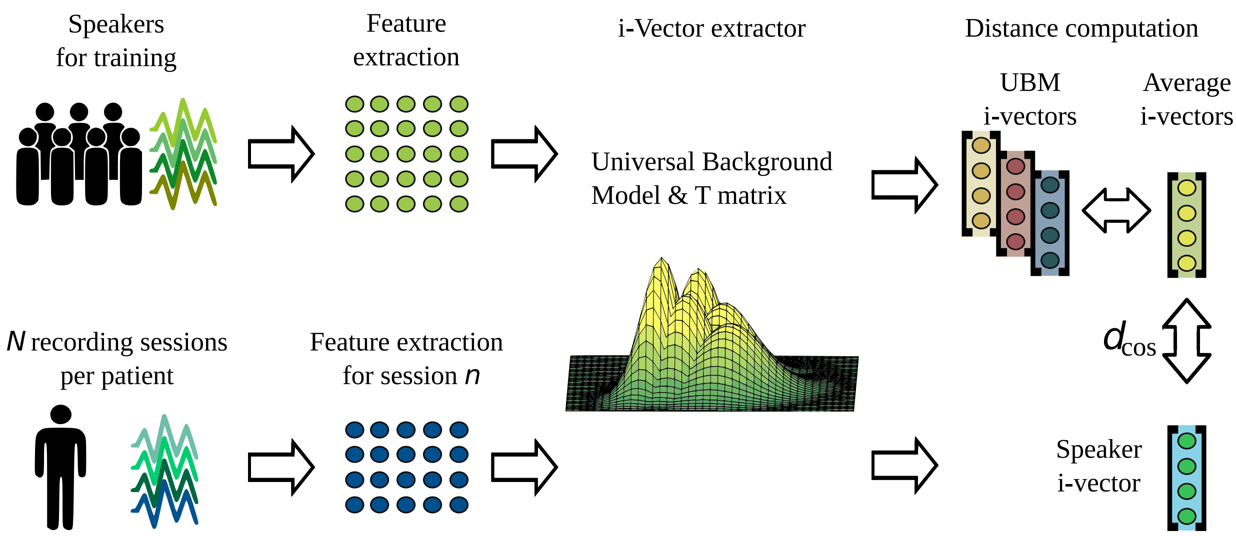


Figure 5: *Speaker modeling. PD progression in N recording sessions per patient:* $n \in \{1, 2, 3, ..., N\}$.

found and assigned as the weight for the seventh speaker which is the person to whom the model is being tested. The procedure is performed for all of the seven speakers in the test set.

2.4. Disease progression

Parkinson's is a progressive disease, thus symptoms severity get worse over the time. According to previous studies, the speech of PD patients is impaired and such an impairment progresses with the disease [6]. The hypothesis is that these variations in the speech of the patients may be reflected in the evaluation performed by the phoniatrician. The goal of the speaker model is to identify changes in the speech of the patient over the time. One way to achieve this aim is to compute the distance between the UBM and the speaker model. Since the distances are estimated considering the same speech recordings evaluated by the phoniatrician, it is expected that the trend of those distances follows the trend of the m–FDA scores. Figure 6 shows a graphical representation of the described situation for the patient 1 in the longitudinal data set (Table 2). The dotted black curve represents the trend of the disease progression for the patient who was evaluated in different sessions, and the gray curve represents the distances computed from the speaker models.

2.5. Non-controlled acoustic conditions

Although the proposed approach seems to be convincing and appropriate for the aforementioned tasks, it is necessary to test its suitability in more realistic conditions. Considering that nowadays most of the people have access to different communication ways, e.g., mobile phones, Skype®, Hangouts®, or landlines, we decided to include all of these options in the experimental setup. The UBM models are trained considering the above mentioned communication ways in order to make the approaches more robust to different acoustic conditions. Nevertheless, there may be a loss of information for particular sets of features. For

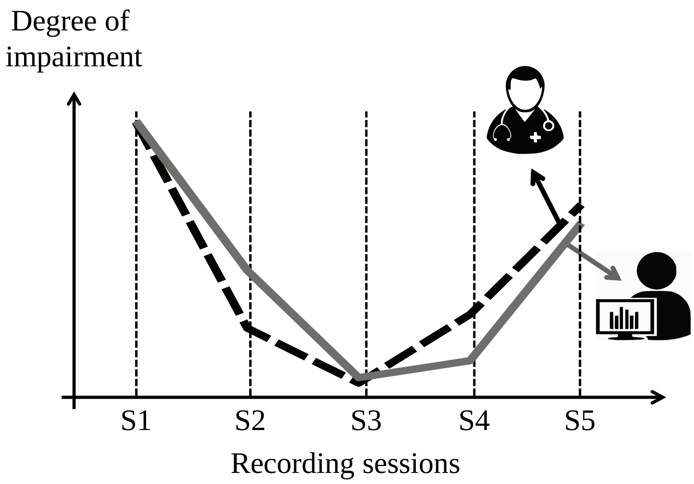


Figure 6: *Graphical representation of the progression of PD for patient 1.*

instance, during a mobile phone call speech signals are sampled at 8 kHz, which limits the computation of the Bark energies to 17 frequency bands. Table 4 indicates the transmission rates of the five communication channels used in the training process.

Channel	Mobile	Landline	Skype®	Hangouts®	Original
Transmission rate (kbps)	6.60 – 23.85	56	6 – 40	6 – 510	256

Table 4: *Transmission rates (kbps) for the five channels considered in this study.*

3. Experiments and results

The original speech signals of the training set were recorded in a sound proof both (PC-GITA database). Then, we re-captured the speech recordings through Skype® calls, Google Hangouts® conversations, landline calls, and mobile phone calls. Each speech aspect (phonation, articulation, and prosody) is modeled and tested considering three different approaches: SVR, GMM–UBM, and i–vectors. Five different models are created, one per communication channel. Spearman's correlation coefficients are used to evaluate the results of the at-home test sets; however, in the longitudinal test set the Pearson's correlation coefficient was used due to the reduced amount of recording sessions. Additionally, the Mean Squared Error (MSE) is computed to evaluate the capability of the speaker models to monitor speech-related problems due to PD.

Experiments with the at-home test set

Table 5 shows the results obtained when the SVR is considered to estimate the m–FDA scores for the at-home test set. Each row corresponds to the Spearman's correlation coefficient between the estimated scores and the real m–FDA obtained with the SVR trained per speech aspect and communication channel. It can be observed that none of the results were satisfactory. The highest correlations were obtained only for patient 1 when the articulation features were considered to train the SVR. This can be likely explained because typically, the SVRs are used to estimate labels, e.g., the dysarthria level, of a group of speakers rather than to monitor each patient individually. The results obtained with the different communication channels indicate that the SVR seems to be not suitable to estimate the dysarthria level when the acoustic conditions are not controlled.

Table 5: Spearman's correlation coefficient (ρ) between the estimated scores and the m–FDA label per patient in the at-home test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average MSE per communication channel.

SVR	Channel	P1	P2	P3	P4	P5	P6	P7	AVG	MSE
Phonation	Original	0.78	-0.06	0.04	0.03	0.00	0.08	0.31	0.17	1.73
	Skype®	-0.24	-0.22	0.22	-0.40	0.20	-0.06	0.18	-0.05	2.03
	Mobile	-0.01	-0.27	-0.11	-0.55	0.14	-0.05	0.18	-0.10	2.14
	Landline	0.06	-0.35	0.03	-0.04	0.04	-0.17	0.40	-0.00	1.96
	Hangouts®	-0.22	-0.41	0.09	-0.52	0.26	-0.18	0.08	-0.13	2.24
Prosody	Original	0.73	-0.00	0.19	0.04	0.11	0.12	0.30	0.21	1.61
	Skype®	0.39	0.24	-0.00	-0.03	0.02	-0.08	0.09	0.09	1.71
	Mobile	-0.20	0.55	-0.49	0.19	-0.18	0.06	0.12	0.01	1.97
	Landline	0.20	0.01	0.11	0.33	-0.33	-0.01	0.00	0.04	1.93
	Hangouts®	0.08	-0.12	-0.55	0.45	-0.46	-0.13	0.03	-0.10	2.19
Articulation	Original	0.49	-0.39	0.03	-0.44	0.71	0.20	-0.00	0.09	1.75
	Skype®	0.43	0.19	-0.07	-0.28	0.65	-0.06	0.24	0.16	1.69
	Mobile	0.28	-0.37	0.01	-0.33	-0.70	-0.21	-0.23	-0.22	2.54
	Landline	0.88	0.08	-0.04	-0.50	0.48	0.51	0.03	0.21	1.47
	Hangouts®	0.55	-0.36	-0.17	0.24	-0.10	-0.49	-0.03	-0.05	2.01

Table 6 shows the Spearman's correlation coefficients between the Bhattacharyya-based similarity measure and the m–FDA for the at-home test set. Each row corresponds to the correlation coefficient obtained with different GMM–UBMs created per speech aspect and communication channel. It can be observed that the highest average correlations are obtained for the GMM–UBMs trained with the articulation features. The results in the last column of Table 6 (AVG) indicate that the average performance of the speaker models per speech aspect is similar in all of the communication channels, indicating that the proposed approach is robust against non-controlled acoustic conditions and communication channels. Although the best results are obtained with the articulation features extracted from the original speech recordings ($\rho = 0.45$), a similar result was obtained for the Skype® calls ($\rho = 0.44$) with the lowest MSE (1.07).

Table 7 shows the Spearman's correlations between the i–vectors-based similarity measure and the m–FDA for the at-home test set. Each row corresponds to the correlation coefficient obtained with different i–vectors created per speech aspect and communication channel. Similar to the GMM–UBM approach, the best results were obtained with the articulation features, particularly, for the original speech recordings and the Skype® calls

Table 6: Spearman's correlation coefficient (ρ) between Bhattacharyya-based similarity measure and m–FDA per patient in the at-home test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average MSE per communication channel.

GMM–UBM	Channel	P1	P2	P3	P4	P5	P6	P7	AVG	MSE
Phonation	Original	0.42	0.12	0.35	0.31	0.63	0.49	0.36	0.38	1.28
	Skype®	0.80	0.50	0.15	0.32	0.26	0.41	0.37	0.40	1.22
	Mobile	0.63	0.28	0.19	0.41	0.17	0.31	0.50	0.36	1.32
	Landline	0.42	0.08	0.35	0.35	0.62	0.39	0.35	0.37	1.39
	Hangouts®	0.72	0.56	0.03	0.53	0.16	0.42	0.23	0.38	1.36
Prosody	Original	0.47	0.66	0.10	0.12	0.38	0.18	0.23	0.31	1.57
	Skype®	0.33	0.16	0.15	0.31	0.42	0.38	0.36	0.30	1.45
	Mobile	0.35	0.19	0.20	0.33	0.42	0.15	0.15	0.26	1.71
	Landline	0.39	0.11	0.31	0.29	0.46	0.40	0.19	0.31	1.58
	Hangouts®	0.29	0.09	0.40	0.06	0.54	0.26	0.39	0.29	1.52
Articulation	Original	0.79	0.05	0.23	0.30	0.79	0.46	0.53	0.45	1.22
	Skype®	0.73	0.12	0.00	0.47	0.83	0.50	0.41	0.44	1.07
	Mobile	0.78	0.00	0.18	0.51	0.56	0.50	0.15	0.38	1.15
	Landline	0.74	0.13	0.21	0.47	0.75	0.48	0.19	0.42	1.18
	Hangouts®	0.76	0.33	0.10	0.01	0.80	0.39	0.45	0.41	1.26

($\rho = 0.41$).

Besides estimating the m-FDA scores considering each speech aspect separately (phonation, articulation, and prosody) and considering that the m–FDA scale includes several items per speech aspect, we wanted to evaluate how much we can improve when those scores are estimated by combining all of the speech aspects. For the case of the GMM–UBM and i–vectors approaches, we combined the information of the three aspects by using Equation 13. For the case of the SVR, the three feature sets are concatenated by using four functionals: mean, standard deviation, skewness, and kurtosis. Table 8 shows the obtained results when the three sets of features are combined. It can be observed that there is an improvement in most of the cases, except for the SVR. The best results are obtained when the speaker models are created using the GMM–UBM approach. This can be explained because in that case the computed distance (the Battacharyya distance) incorporates several characteristics of the model, i.e., mean vectors, covariance matrices, and the weights of the Gaussian mixture. For the case of the i–vectors, the distance between each speaker and the UBM is computed using the dot product between the i–vector of the speaker who is being monitored and the average i–vector computed over all of the speakers in the UBM. This procedure may cause a loss of information about the variability of the models. The results obtained with the SVR are not satisfactory because the regressor is trained to estimate the dysarthria level of a group of speakers but not a specific speaker. One hypothesis is that this results could be improved if an SVR is trained per patient. However, more recording sessions per speaker are necessary to obtain more conclusive results. Regarding the results obtained per channel, the best are obtained with the Skype® calls ($\rho = 0.55$, MSE$= 0.89$) for the GMM–UBM.

Table 7: Spearman's correlation coefficient (ρ) between the i–vector-based similarity measure and the m–FDA score per patient in the at-home test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average MSE per communication channel.

i–vectors	Channel	P1	P2	P3	P4	P5	P6	P7	AVG	MSE
Phonation	Original	0.56	0.02	0.28	0.02	0.04	0.04	0.26	0.17	1.81
	Skype®	0.06	0.27	0.20	0.31	0.11	0.12	0.14	0.17	1.83
	Mobile	0.59	0.06	0.12	0.01	0.01	0.05	0.11	0.14	1.83
	Landline	0.34	0.57	0.13	0.53	0.27	0.12	0.16	0.30	1.76
	Hangouts®	0.28	0.01	0.30	0.04	0.36	0.23	0.14	0.19	1.81
Prosody	Original	0.61	0.36	0.16	0.25	0.24	0.31	0.30	0.32	1.28
	Skype®	0.15	0.05	0.48	0.18	0.41	0.14	0.11	0.22	1.43
	Mobile	0.73	0.10	0.34	0.02	0.27	0.38	0.15	0.28	1.43
	Landline	0.45	0.33	0.02	0.26	0.30	0.37	0.27	0.29	1.50
	Hangouts®	0.53	0.49	0.11	0.39	0.43	0.08	0.22	0.32	1.23
Articulation	Original	0.64	0.71	0.22	0.30	0.14	0.53	0.31	0.41	1.18
	Skype®	0.33	0.05	0.39	0.51	0.74	0.35	0.49	0.41	1.27
	Mobile	0.49	0.21	0.27	0.31	0.77	0.34	0.28	0.38	1.22
	Landline	0.62	0.08	0.11	0.35	0.72	0.29	0.00	0.31	1.27
	Hangouts®	0.28	0.00	0.51	0.48	0.66	0.27	0.05	0.32	1.28

This can be explained considering the preprocessing stage performed

This result is very promising because it opens the possibility of developing computer-aided tools to monitor symptoms developed by PD patients using conventional Skype® calls.

Figure 7 displays curves with the comparison of the estimated m–FDA scores (red lines) and the real labels assigned by the phoniatrician (black lines). The x-axis represents the recording session. For the red lines, the y-axis represents the normalized values of the multi-aspect coefficient ξ, estimated using the GMM–UBM approach combining phonation, articulation, and prosody distances. For the black lines, the y-axis represents the normalized original m–FDA scores. The normalization is performed using the z-score approach only for displaying purposes, i.e., to depict comparable curves in the same figure. The distances computed from each speaker model represent the progression of the dysarthria level due to the disease progression. Note that the trend of the estimated scores follows the trend of the dysarthria level in most of the cases. According to Figure 7, the proposed approach seems to be suitable to monitor the progression of the dysarthria level in PD patients; however, further research is required to include more patients and recording sessions, and also to consider the possible variability introduced by the medication intake.

Experiments with the longitudinal test set – Dysarthria level assessment

Table 9 shows the results obtained with the SVR when the dysarthria levels of the speakers in the longitudinal test set are considered. As in the previous experiments with the at-home test set, this result indicates that the SVR approach it is not suitable to monitor

Table 8: Spearman's correlation coefficient (ρ) between the multi-aspect coefficient ξ and m–FDA per patient in the at-home test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average Mean Squared Error.

Model	Channel	P1	P2	P3	P4	P5	P6	P7	AVG	MSE
SVR	Original	0.46	-0.49	0.18	-0.35	-0.01	0.26	0.12	0.02	1.85
	Skype®	0.39	0.21	-0.20	-0.29	0.61	-0.07	0.20	0.12	1.72
	Mobile	0.82	-0.01	-0.09	-0.37	0.37	0.10	0.37	0.17	1.99
	Landline	-0.08	-0.03	0.16	-0.15	0.07	0.23	-0.12	0.01	1.47
	Hangouts®	0.30	-0.15	-0.29	-0.18	0.05	-0.00	-0.06	-0.05	2.14
GMM–UBM	Original	0.62	0.44	0.22	0.31	0.86	0.44	0.39	0.47	1.07
	Skype®	0.76	0.54	0.19	0.46	0.86	0.48	0.54	0.55	0.89
	Mobile	0.61	0.25	0.24	0.67	0.77	0.29	0.26	0.44	1.26
	Landline	0.73	0.57	0.06	0.40	0.87	0.56	0.47	0.51	1.00
	Hangouts®	0.70	0.49	0.23	0.50	0.45	0.66	0.30	0.48	1.22
i–vectors	Original	0.63	0.53	0.12	0.46	0.14	0.48	0.30	0.38	1.14
	Skype®	0.26	0.00	0.33	0.67	0.58	0.34	0.61	0.40	1.26
	Mobile	0.54	0.24	0.36	0.41	0.77	0.31	0.27	0.41	1.13
	Landline	0.68	0.07	0.22	0.63	0.46	0.49	0.23	0.38	1.40
	Hangouts®	0.59	0.32	0.28	0.45	0.66	0.39	0.34	0.43	1.04

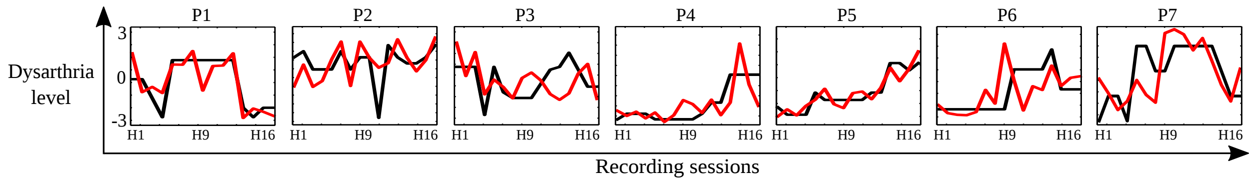


Figure 7: *Curves of the dysarthria level per patient (***Pi***) in the at-home test set. Comparison of the m-FDA scores estimated using GMM–UBM with the Skype® recordings (red lines) and the original m–FDA values assigned by the phoniatricians (black lines) for the at-home test set.*

the progression of PD patients individually.

Table 10 shows the results when the GMM–UBM was created with phonation, prosody, and articulation features, separately. It can be observed that the average performance per channel is similar for the three speech aspects. Note that, the results in the longitudinal test set are better than those obtained with the at-home test set. This can be likely explained because the at-home test is more challenging. During the collection of the at-home samples the speech of the patient may change due to several factors like medication intake, mood, tiredness, etc. The estimation of the dysarthria level in the longitudinal test set is like the analysis of a "picture" taken every six months approximately, while the study of the at-home test set is like a zoom-in of the dysarthria level of the patient, i.e., every two hours, once per month during four months.

Table 11 shows the results for the i–vectors approach. Similar to the GMM–UBM, there is an improvement in the performance of the models with respect to the results in the at-home test set. As in the previous experiments, the best results are obtained with the

Table 9: Pearson's correlation coefficient (r) between the estimated scores and the m–FDA score per patient in the longitudinal test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average Mean Squared Error.

SVR	**Channel**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**	**MSE**
Phonation	Original	-0.40	0.32	0.00	0.70	0.36	0.50	-0.40	0.15	1.75
	Skype®	0.10	-0.32	0.30	-0.70	-0.46	-0.50	-0.50	-0.30	1.84
	Mobile	-0.10	-0.63	0.70	-0.40	-0.62	-0.20	-0.50	-0.25	1.92
	Landline	0.50	0.32	-0.30	-0.20	-0.46	-0.90	-0.60	-0.23	2.02
	Hangouts®	0.20	0.63	0.10	-0.70	-0.62	0.50	-0.50	-0.06	2.04
Prosody	Original	-0.80	-0.63	-0.50	0.70	0.62	-0.70	0.60	-0.10	2.09
	Skype®	-0.13	-0.52	-0.20	-0.02	0.49	-0.41	0.72	-0.01	2.02
	Mobile	-0.50	0.95	-0.60	-0.30	-0.62	-0.60	0.10	-0.22	2.32
	Landline	0.10	0.32	-0.10	1.00	-0.82	-0.20	0.00	0.04	2.19
	Hangouts®	-0.05	-0.28	0.49	-0.69	-0.48	0.04	-0.01	-0.14	2.28
Articulation	Original	-0.70	0.32	-1.00	0.30	-0.72	-0.20	0.40	-0.23	2.53
	Skype®	-0.20	-0.32	-0.70	0.10	-0.62	-0.10	-0.60	-0.35	2.61
	Mobile	0.70	0.32	0.00	-0.10	0.36	0.30	0.80	0.34	1.14
	Landline	-0.30	-0.32	-0.90	-0.50	0.31	-0.20	0.10	-0.26	2.30
	Hangouts®	-0.40	-0.63	-0.50	-0.30	0.10	0.30	-0.50	-0.28	2.47

Table 10: Pearson's correlation coefficient (r) between the Bhattacharyya-based similarity measure and the m–FDA score per patient in the longitudinal test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average Mean Squared Error.

GMM–UBM	**Channel**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**	**MSE**
Phonation	Original	0.84	0.50	0.81	0.15	0.78	0.69	0.30	0.58	0.87
	Skype®	0.51	0.20	0.26	0.78	0.92	0.38	0.35	0.49	1.10
	Mobile	0.43	0.73	0.37	0.97	0.34	0.53	0.48	0.55	0.85
	Landline	0.61	0.51	0.40	0.43	0.92	0.30	0.46	0.52	0.90
	Hangouts®	0.86	0.11	0.44	0.57	0.62	0.31	0.38	0.47	1.03
Prosody	Original	0.10	0.65	0.31	0.34	0.66	0.91	0.93	0.56	0.90
	Skype®	0.80	0.99	0.17	0.40	0.35	0.53	0.55	0.54	1.06
	Mobile	0.85	0.54	0.63	0.40	0.30	0.73	0.31	0.54	0.92
	Landline	0.87	0.85	0.32	0.89	0.24	0.19	0.41	0.54	0.92
	Hangouts®	0.90	0.92	0.48	0.25	0.64	0.01	0.67	0.55	0.87
Articulation	Original	0.46	0.62	0.23	0.48	0.93	0.42	0.69	0.55	1.08
	Skype®	0.71	0.25	0.71	0.42	0.23	0.63	0.64	0.51	0.91
	Mobile	0.39	0.84	0.04	0.69	0.39	0.68	0.90	0.56	0.94
	Landline	0.36	0.77	0.24	0.25	0.94	0.94	0.72	0.60	0.78
	Hangouts®	0.63	0.46	0.90	0.56	0.92	0.67	0.12	0.61	0.81

articulation features.

Note that the GMM–UBM approach is better than the others when phonation and prosody features are considered, while the i–vectors approach is better when the speech signals are modeled with the articulation features. Although the difference in the results obtained with these two approaches using the articulation and prosody features is not very big, the use of i–vectors could be more convenient when the recording conditions (acoustic environment, microphone, and sound card) are not controlled.

Table 11: Pearson's correlation coefficient (r) between the dot product-based similarity measure and the m–FDA score per patient in the longitudinal test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average Mean Squared Error.

i–vectors	**Channel**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**	**MSE**
Phonation	Original	0.69	0.40	0.24	0.43	0.42	0.11	0.71	0.43	1.14
	Skype®	0.11	0.58	0.58	0.43	0.36	0.44	0.15	0.38	1.33
	Mobile	0.84	0.31	0.39	0.36	0.52	0.09	0.12	0.38	1.32
	Landline	0.80	0.20	0.33	0.09	0.42	0.35	0.93	0.45	1.22
	Hangouts®	0.79	0.08	0.32	0.36	0.60	0.95	0.42	0.50	1.54
Prosody	Original	0.62	0.46	0.91	0.87	0.96	0.04	0.08	0.56	0.88
	Skype®	0.81	0.55	0.64	0.35	0.81	0.04	0.05	0.46	1.07
	Mobile	0.77	0.49	0.87	0.84	0.47	0.27	0.04	0.54	0.92
	Landline	0.29	0.75	0.56	0.62	0.78	0.32	0.17	0.50	1.00
	Hangouts®	0.17	0.04	0.53	0.82	0.93	0.18	0.63	0.47	1.06
Articulation	Original	0.80	0.89	0.97	0.55	0.61	0.33	0.06	0.60	0.80
	Skype®	0.49	0.39	0.75	0.41	0.79	0.94	0.76	0.65	0.70
	Mobile	0.49	0.19	0.13	0.98	0.98	0.72	0.89	0.63	0.75
	Landline	0.52	0.25	0.27	0.97	0.73	0.67	0.91	0.62	0.76
	Hangouts®	0.66	0.78	0.41	0.87	0.70	0.20	0.43	0.58	0.85

Besides the analysis of each speech aspect separately, it is also interesting to evaluate the convenience of their combination. In order to do that we use the multi-aspect coefficient ξ introduced in Equation 13. Table 12 shows the Pearson's correlation coefficients between ξ and the original m–FDA scores. Note that similar to the at-home test set, there is an improvement when the distances are combined. In this case, the higher correlation is achieved with the i–vectors trained considering the original speech recordings ($r = 0.77$, MSE$= 0.47$). These results are better than those obtained in the experiments with the at-home test set. As it was mentioned above, this is because the time between sessions in the at-home test set is shorter than the time between sessions in the longitudinal test set, thus the disease progression can be more accurately modeled in the longitudinal test set than in the at-home test set. The acoustic conditions between recording sessions in the longitudinal set could have changed more than between the recording sessions in the at-home set. This could explain why the i–vectors show to be more accurate than the GMM–UBM in the longitudinal recordings.

Table 12: Pearson's correlation coefficient (ρ) between the multi-aspect coefficient ξ and m–FDA per patient in the longitudinal test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average Mean Squared Error.

Model	Channel	P1	P2	P3	P4	P5	P6	P7	AVG	MSE
SVR	Original	-0.74	-0.57	-0.95	0.46	-0.50	-0.29	0.13	-0.35	2.70
	Skype®	0.89	-0.94	-0.63	-0.21	-0.54	-0.09	-0.19	-0.24	2.49
	Mobile	-0.08	0.52	0.26	-0.64	0.30	0.36	-0.42	0.04	1.91
	Landline	-0.57	-0.02	-0.79	0.21	-0.18	-0.56	0.21	-0.24	2.49
	Hangouts®	-0.50	0.23	-0.48	-0.91	-0.07	0.43	-0.38	-0.24	2.48
GMM–UBM	Original	0.85	0.76	0.74	0.26	0.95	0.85	0.36	0.68	0.64
	Skype®	0.80	0.55	0.55	0.58	0.29	0.65	0.70	0.59	0.82
	Mobile	0.55	0.79	0.16	0.75	0.79	0.75	0.76	0.65	0.79
	Landline	0.75	0.90	0.40	0.53	0.85	0.91	0.63	0.71	0.58
	Hangouts®	0.82	0.60	0.89	0.51	0.86	0.63	0.15	0.64	0.73
i–vectors	Original	0.81	0.94	0.88	0.65	0.96	0.39	0.75	0.77	0.47
	Skype®	0.73	0.80	0.96	0.53	0.87	0.82	0.50	0.74	0.52
	Mobile	0.68	0.43	0.44	0.97	0.88	0.81	0.55	0.68	0.64
	Landline	0.51	0.24	0.34	0.85	0.79	0.60	0.81	0.59	0.81
	Hangouts®	0.49	0.47	0.53	0.89	0.84	0.13	0.67	0.54	0.93

Figure 8 shows the trends obtained when the m–FDA scores are estimated with the coefficient ξ when its aspects are modeled using the i–vectors approach and the median of the original m-FDA scores assigned by the phoniatricians. The x-axis indicates the five recording sessions of the longitudinal dataset. This figure displays only the best results which are based on the original recordings, i.e., without any transmission over telephone or Internet channels. Note that the i–vectors approach shows to be accurate and robust to estimate the trend of the dysarthria level in PD patients, which motivates us to continue exploring these methods to perform the automatic and unobtrusive monitoring of PD.

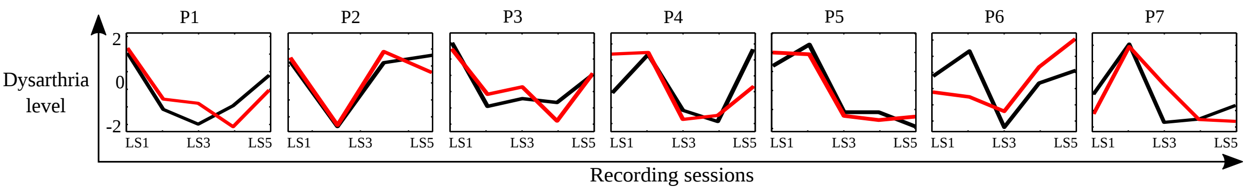


Figure 8: *Curves of the dysarthria level per patient (***Pi***) in the longitudinal test set. Comparison of the m-FDA scores estimated using i–vectors with the original recordings (red lines) and the original m–FDA values assigned by the phoniatricians (black lines).*

Experiments with the longitudinal test set – Neurological evaluation

Besides the evaluation of the dysarthria level, the neurological state of the patients in the longitudinal test set is considered (it was not possible to perform neurological evaluations during the at-home recordings). In this case, the Pearson's correlation (r) is estimated between the multi-aspect coefficient ξ and the MDS-UPDRS-III scores assigned by neurologist. The influence of the five communication channels is also evaluated. The results are displayed in Table 13.

Table 13: Pearson's correlation coefficients (r) estimated between ξ calculated using i–vectors and MDS-UPDRS-III per patient in the longitudinal test set (**Pi**). **AVG**: Average correlation per communication channel. **MSE**: Average Mean Squared Error.

Channel	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**AVG**	**MSE**
Original	0.31	-0.85	0.93	0.40	-0.35	0.65	0.08	0.17	1.38
Skype®	0.70	0.99	0.93	0.54	0.28	-0.03	0.41	0.55	0.89
Mobile	0.57	-0.77	0.94	-0.03	-0.57	0.63	-0.98	-0.03	1.98
Landline	0.82	0.20	0.69	-0.37	0.25	-0.33	-0.99	0.04	1.68
Hangouts®	0.88	0.28	0.49	0.42	-0.15	0.05	-0.77	0.17	1.36

Note that these results are not as good as those obtained when evaluating the dysarthria level. This behavior was expected because the m–FDA scale was designed to assess speech impairments related with dysarthria, while the MDS-UPDRS-III is used to assess general motor symptoms in PD patients. Although the correlations are not high, the results for P1, P2, P3, and P7 indicate that, to some extent, the speech impairments modeled by the proposed approach have impact in the general motor state of the patients. Additionally, note that the trends of the curves are similar in most of the cases. Although we could claim that the proposed approach is promising to evaluate the neurological state of PD patients, the main conclusion is that the use of multi-modal approaches are required in order to obtain more accurate and reliable results.

Figure 9 displays the trend of the estimated MDS-UPDRS-III and the original labels assigned by the neurologist. Only the curves of the best result (with the recordings captured using Skype® calls) are displayed. As in the previous cases, the values are z-score normalized to allow direct comparison between the trends of the two curves. Red lines represent the estimated MDS-UPDRS-III scores and the black lines indicate the original MDS-UPDRS-III scores. The x-axis includes all of the five recording sessions of the longitudinal test set. Note that in this case the results are not satisfactory in all of the cases, the trends coincide in four of the seven patients (P1, P2, P3, and P4). We expect that these results can be significantly improved when considering multi-modal approaches.

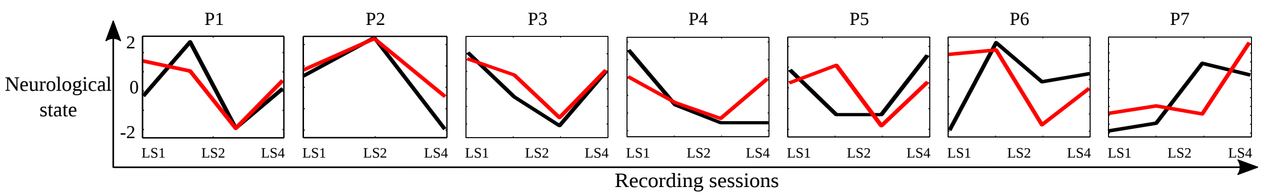


Figure 9: *Curves of the neurological level per patient (***Pi***). Comparison of the MDS-UPDRS-III scores estimated using i–vectors with the recordings of the Skype® calls (red lines) and original MDS-UPDRS-III values assigned by the neurologist expert (black lines).*

4. Conclusions

This paper presents a methodology for the automatic monitoring of speech disorders developed by PD patients. The neurological state and dysarthria level of the patients are evaluated. The proposed system is based on individual speaker models which are created for each patient. Two different models are evaluated, the classical GMM–UBM and the i–vectors approach. These two methods are compared with respect to a baseline found with a traditional Support Vector Regressor. Different speech aspects (phonation, articulation, and prosody) are considered to model the speech signals and a multi-aspect coefficient is proposed with the aim of incorporating information from all of these speech aspects into a single measure. Two different scenarios are considered to assess a set with seven PD patients: (1) the longitudinal test set which consists of speech recordings captured in five recordings sessions distributed from 2012 to 2016, and (2) the at-home test set which consists of speech recordings captured in the home of the same seven patients during 4 months (once day per month, four times per day). The UBM is trained with speech recordings captured in a sound-proof booth, i.e., with controlled acoustic conditions and a professional audio-setting. With the aim of evaluating the suitability of the proposed approaches and the possibility of extending this kind of systems to remotely assess the speech of the patients, a total of five different communication channels (sound-proof booth, Skype®, Hangouts®, mobile phone, and land-line) are considered to train and test the system.

According to the results, it is possible to track the dysarthria level of PD patients when the three speech aspects are combined. It seems like Skype® is the best alternative to perform the analyses; however, the difference with respect to the other channels is not high, which indicates that the proposed approaches are promising for the continuous and unobtrusive monitoring of PD progression.

The results obtained when assessing the neurological state of the patients are not satisfactory. This can be explained due to the fact that the MDS-UPDRS-III scale comprises a total of 33 items to evaluate the general motor capability of the patients but the speech is only considered in one of the items. Further research is required in order to obtain more conclusive and accurate results. We think that the inclusion of information from more bio-signals, i.e., multi-modal systems, will lead to more accurate, stable, and conclusive results. We are currently working on the construction of a dataset with different sensor-data and we expect to be able to improve the current results in the near future. Besides the multi-modal modeling, the study of the variability in the speech of PD patients due to the medication intake will be considered in future works.

Acknowledgments

This project was funded by CODI at Universidad de Antioquia (grants # PRV16-2-01 and 2015-7683). The work has received also funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement no. 766287. Tomás Arias-Vergara is under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS. The authors would like to thank all of the patients and collaborators from Fundalianza Parkinson Colombia. Without their support and contribution it would not be possible to address this research.

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Additionally, the procedures were approved by the Ethics Committee of Universidad de Antioquia and Clínica Noel, in Medellín, Colombia.

Informed consent

Informed consent was obtained from all of the persons who participated in this study.

References

[1] O. Hornykiewicz, Biochemical aspects of Parkinson's disease, Neurology 51 (2) (1998) S2–S9.

[2] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, S. Gates, Speech impairment in a large sample of patients with parkinson's disease, Behavioural neurology 11 (3) (1999) 131–137.

[3] F. L. Darley, A. E. Aronson, J. R. Brown, Differential diagnostic patterns of dysarthria, Journal of Speech, Language, and Hearing Research 12 (2) (1969) 246–269.

[4] D. G. Theodoros, G. Constantinescu, T. G. Russell, E. C. Ward, S. J. Wilson, R. Wootton, Treating the speech disorder in Parkinson's disease online, Journal of Telemedicine and Telecare 12 (suppl 3) (2006) 88–91.

[5] A. Tsanas, M. Little, P. E. McSharry, L. Ramig, Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests, IEEE Transactions on Biomedical Engineering 57 (4) (2010) 884–893.

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

[7] J. R. Orozco-Arroyave, F. Hönig, et al., Automatic detection of parkinson's disease in running speech spoken in three different languages, The Journal of the Acoustical Society of America 139 (1) (2016) 481–500.

[8] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease and others, The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations, Movement disorders: official journal of the Movement Disorder Society 18 (7) (2003) 738.

[9] 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 23 (15) (2008) 2129–2170.

[10] C. G. Goetz, et al., Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: status and recommendations the Movement Disorder Society Task Force on rating scales for Parkinson's disease, Movement disorders 19 (9) (2004) 1020–1028.

[11] P. M. Enderby, R. Palmer, FDA-2: Frenchay Dysarthria Assessment: Examiner's Manual, Pro-ed, 2008.

[12] M. Cernak, J. Orozco-Arroyave, F. Rudzicz, H. Chirstensen, J. Vásquez-Correa, E. Nöth, Characterization of voice quality of Parkinson's disease using differential phonological posterior features, Computer Speech and Language 46 (2017) 196–208.

[13] M. Asgari, I. Shafran, Extracting Cues from Speech for Predicting Severity of Parkinson's Disease, in: IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2010, pp. 462–467.

[14] A. Bayestehtashk, M. Asgari, I. Shafran, J. McNames, Fully automated assessment of the severity of Parkinson's disease from speech, Computer Speech and Language 29 (1) (2015) 172–185.

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

[16] B. Schuller, S. Steidl, et al., The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition, in: Proceeding of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015, pp. 478–482.

[17] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, E. Nöth, New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease, in: Proceedings of the 9th International Conference on Language Resources and Evaluation, 2014, pp. 342–347.

[18] T. Grósz, R. Busa-Fekete, G. Gosztolya, L. Tóth, Assessing the Degree of Nativeness and Parkinson's Condition Using Gaussian Processes and Deep Rectifier Neural Networks, in: Proceeding of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015, pp. 919–923.

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

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

[21] P. Gómez-Vilda, M. C. Vicente-Torcal, J. M. Ferrández-Vicente, A. Álvarez-Marquina, V. Rodellar-Biarge, V. Nieto-Lluis, R. Martínez-Olalla, Parkinson's Disease Monitoring from Phonation Biomechanics, Lecture Notes in Computer Science 9107 (2015) 238–248.

[22] P. Gómez-Vilda, A. Álvarez-Marquina, V. Rodellar-Biarge, Monitoring Parkinson's Disease from phonation improvement by Log Likelihood Ratios, in: Bioinspired Intelligence (IWOBI), Fourth International Work Conference on, 2015, pp. 105–110.

[23] T. Arias-Vergara, J. Vásquez-Correa, J. Orozco-Arroyave, J. Vargas-Bonilla, E. Nöth, Parkinson's disease progression assessment from speech using GMM-UBM, in: Proceeding of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2016, pp. 1933–1937.

[24] N. García, J. Orozco-Arroyave, L. D'Haro, N. Dehak, E. Nöth, Evaluation of the neurological state of people with Parkinson's disease using i-vectors, in: Proceeding of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2017, pp. 299–303.

[25] N. García, J. Vásquez-Correa, J. R. Orozco-Arroyave, N. Dehak, E. Nöth, Language independent assessment of motor impairments of patients with parkinson's disease using i-vectors, Lecture Notes in Computer Science 10415 (2017) 147–155.

[26] A. Tsanas, M. Little, P. E. McSharry, J. Spielman, L. O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Transactions on Biomedical Engineering 59 (5) (2012) 1264–1271.

[27] J. C. Vásquez-Correa, J. Serra, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, E. Nöth, Effect of Acoustic Conditions on Algorithms to Detect Parkinson's Disease from Speech, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)), 2017, pp. 5065–5069.

[28] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, et al., Neurospeech: An open-source software for Parkinson's speech analysis, Digital Signal Processing (In press) (2017) 1–15.

[29] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londoño, E. Nöth, Automatic Detection of Parkinson's Disease from Continuous Speech Recorded in Non-Controlled Noise Conditions, in: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015, pp. 105–109.

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

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

[32] N. Dehak, P. Dumouchel, P. Kenny, Modeling prosodic features with joint factor analysis for speaker

verification, IEEE Transactions on Audio, Speech, and Language Processing 15 (7) (2007) 2095–2103.

[33] J. I. Godino-Llorente, P. Gomez-Vilda, M. Blanco-Velasco, Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters, IEEE Transactions on Biomedical Engineering 53 (10) (2006) 1943–1953.

[34] E. Zwicker, E. Terhardt, Analytical expressions for critical-band rate and critical bandwidth as a function of frequency, Journal of the Acoustical Society of America 68 (5) (1980) 1523–1525.

[35] H. Hermansky, B. Hanson, H. Wakita, Perceptually based linear predictive analysis of speech, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1985.

[36] J. O. Smith, J. S. Abel, Bark and ERB bilinear transforms, IEEE Transactions on Speech and Audio Processing 7 (6) (1999) 697–708.

[37] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital signal processing 10 (1) (2000) 19–41.

[38] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 19 (4) (2011) 788–798.

[39] N. Dehak, Najim, Discriminative and generative approaches for long- and short-term speaker characteristics modeling application to speaker verification., Library and Archives Canada, 2010.

[40] C. H. You, K. A. Lee, H. Li, GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition, IEEE Transactions on Audio, Speech, and Language Processing 18 (6) (2010) 1300–1312.

[41] J. C. Gower, P. Legendre, Metric and Euclidean properties of dissimilarity coefficients, Journal of classification 3 (1) (1986) 5–48.