Applied Soft Computing 62 (2018) 649–666

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease

Laureano Moro-Velázquez[a,*], Jorge Andrés Gómez-García[a], Juan Ignacio Godino-Llorente[a], Jesús Villalba[b], Juan Rafael Orozco-Arroyave[c,d], Najim Dehak[b]

[a] Center for Biomedical Technology, Universidad Politécnica de Madrid, Campus Montegancedo, 28223 Pozuelo de Alarcón, Madrid, Spain
[b] ECE, Johns Hopkins University, Homewood Campus, 21218 Baltimore, MD, USA
[c] Universidad de Antioquia, 050010 Medellín, Colombia
[d] Pattern Recognition Lab, University of Erlangen, 91058 Erlangen, Germany

ARTICLE INFO

Article history:
Received 28 December 2016
Received in revised form 18 August 2017
Accepted 1 November 2017

Keywords:
Parkinson's Disease
Speech
GMM-UBM
iVectors
Speaker recognition techniques

ABSTRACT

The diagnosis of Parkinson's Disease is a challenging task which might be supported by new tools to objectively evaluate the presence of deviations in patient's motor capabilities.

To this respect, the dysarthric nature of patient's speech has been exploited in several works to detect the presence of this disease, but none of them has deeply studied the use of state-of-the-art speaker recognition techniques for this task.

In this paper, two classification schemes (GMM-UBM and *i-Vectors*-GPLDA) are employed separately with several parameterization techniques, namely PLP, MFCC and LPC. Additionally, the influence of the kinetic changes, described by their derivatives, is analysed.

With the proposed methodology, an accuracy of 87% with an AUC of 0.93 is obtained in the optimal configuration. These results are comparable to those obtained in other works employing speech for Parkinson's Disease detection and confirm that the selected speaker recognition techniques are a solid baseline to compare with future works. Results suggest that Rasta-PLP is the most reliable parameterization for the proposed task among all the tested features while the two employed classification schemes perform similarly. Additionally, results confirm that kinetic changes provide a substantial performance improvement in Parkinson's Disease automatic detection systems and should be considered in the future.

1. Introduction

The second most prevalent neurodegenerative disease, Parkinson's Disease (PD), is usually diagnosed on the basis of the observation of motor *cardinal signs* [1] and other non-motor indicators (physiological and cognitive manifestations) which are employed in the *clinical diagnosis*. Despite neuropathological diagnosis during autopsy is considered as the gold standard, some studies demonstrate that following the usual clinical diagnosis criteria it is possible to obtain 90% of accuracy in final judgement, but the average detection time to reach this accuracy is 2.9 years [2]. To monitor the progress of the disease, specialists often employ the Unified Parkinson's Disease Rating Scale (UPDRS)[1] [3] or the Hoehn and Yahr (H&Y) scale[2] [4] which include objective and non-objective assessments. In this regard, new technologies could accelerate the diagnosis process and provide a more objective monitoring of the affection.

Since PD affects the coordination of movements, it is reasonable to hypothesize that the assessment of the patients' performance during a complex motor task might be employed for diagnosis purposes. Speech production, an ability that is almost universal, might be affected by PD since it involves complex and very precise movements, but on spite of being good candidate for PD detection and evaluation, its capabilities have not been deeply exploited yet.

* Corresponding author.
E-mail address: laureano.moro@upm.es (L. Moro-Velázquez).

[1] UPRDS comprises four main parts: I, mentation, behaviour and mood; II, activities of daily living; III motor; IV complications. The rating of this scale is obtained through interviews with the patient and clinical observations. The best possible score for each part is 0 whereas the worst one depends on the part. The global UPDRS value can range between 0 and 147, where the larger the value, the higher the affection of the disease.

[2] The H&Y scale comprises several levels whose values can range from 1 to 5 in which 1 implies that the patient has low or no functional disabilities and 5 that patient is totally dependent.

https://doi.org/10.1016/j.asoc.2017.11.001
1568-4946/

It is well known that the neurodegenerative processes associated to the disorder cause *hypokinetic dysarthria*, thus producing a reduction of loudness and articulation amplitude, slowing down the speech sometimes and principally reducing intelligibility [5–9]. Literature evidences the influence of PD on speech from early to advanced stages although it is mainly perceived in mild to advanced phases [10–12]. In this regard, it is expected that new methods of automatic assessment employing voice and speech can be used to detect the signs that are not perceived in the first stages of the disease but which could provide relevant information.

There are several studies and approaches using speech or voice to find biomarkers of the presence of PD or to assess its severity. Most of the literature can be divided into four groups depending on the analysed aspect: *phonatory*, *articulatory*, *prosodic* and *linguistic*. The *phonatory* studies are related to the glottal source and resonant structures of the vocal tract. Works based on *articulatory* and *prosodic* aspects are more abundant and diverse as there exist more analysis possibilities and since the influences of PD in articulation and prosody seem to be more evident [12]. The works within these groups are based on syllable rate analysis, or the processing of certain segments of the speech to obtain indexes correlated with the disease. Concerning the *prosodic* works, studies are mainly focused in the paralinguistic features such as pitch variation or the manifestation of emotions among others [13–17]. Finally, the studies related to deviations in the *linguistic* domain examine the vocabulary, phrase construction and the existence of word repetitions. Some representative works within this group are found in [18–21]. The speech material used in each case is a differentiating factor of the four groups. In the phonatory analysis, the most advisable acoustic material is sustained vowels while in the other three groups, running speech is needed. Specifically, in articulatory analysis, diadochokinetic (DDK)[3] speech can be valuable in addition to the other running speech materials such as spontaneous speech or reading text.

The present study can be framed into both the *articulatory* and the *phonatory* groups attending to the type of analysis that is employed and the acoustic material used.

Going into detail about some *articulatory* relevant works, studies as [23] indicate that speech processing can produce powerful indicators of imprecise consonant articulation in PD-related dysarthria. Authors perform an analysis of DDK tasks (/pa-ta-ka/) in a database of 24 PD patients and 22 controls, providing 88% of efficiency on separating PD from controls. In this study all the utterances are subdivided automatically into different representative segments to analyse articulation. Only 13 features are obtained by performing measurements on these segments, each feature describing a different articulatory trait of speech. Its main drawbacks are the use of a small database which is sex unbalanced and the use of only DDK utterances, limiting the possible articulatory combinations. Other works such as [24] employ frequency features, namely Mel Frequency Cepstrum Coefficients (MFCC) and Band Bark Energies (BBE) from running speech, and other features obtained after the segmentation of specific regions, providing good results with three corpora. However, in this case the results are too optimistic due to an over-fitting of the model, since it was optimized during training.

Equally, there are *articulatory* studies more focused on the fluctuations of the voice onset, offset and break segments during running speech, which are considered to be crucial in the evaluation of voice quality. For instance, in [25,26] it is evidenced that the parkinsonian speech has lower values of relative fundamental frequency, which is the ratio between the fundamental frequency in the cycles of a vowel before or after a voiceless consonant and the typical fundamental frequency during the utterance. The main drawback of these two works is that the databases are unbalanced in sex, which could bias some conclusions. Other studies perform the tracking of vowel formants and VSA during articulation, including onset and offset, with heterogeneous results [27–29]. As formants reflect the position of the tongue, a reduction of the articulation ranges could subsequently limit the frequency ranges of the formants. In [30], a comparison of PD detection techniques is performed using the acoustic material extracted from sustained vowels, sentence repetitions, reading passages and monologues. An accuracy of 80% is achieved using vowels extracted from monologues, providing enhanced results compared to utilizing sustained vowels. The main drawback of this study is the use of a small and unbalanced database (20 patients and 15 controls).

In any case, these and many other works such as [8,9,31,32] evidence that articulation perturbations introduced by PD can provide reliable information about the presence of the disorder.

Respecting the *phonatory* works, sustained vowels are expected to generate simpler acoustic structures that might be easier to analyse. Some works demonstrate that it is possible to detect the influence of PD on the vocal folds vibration by reason of the presence of noise and other perturbations caused by incomplete closure [33], abnormal phase closure and phase asymmetry or vocal tremor [34]. Likewise, some works like [35–37] use dysphonia measures including noise or frequency and amplitude perturbations to assess the severity of PD in telemonitoring scenarios achieving good results. A major drawback, though, is that recordings are done using portable and different equipments introducing noise and variability in the databases which could bias the system.

Finally, other authors employ a combination of techniques as in [38]. In this case, phonatory, prosodic and articulatory features are used jointly, providing results of 80% of accuracy in PD detection.

Although there are many approaches using voice and speech as acoustic materials to detect and assess PD, as far as the authors of this study know, no work has analysed thoughtfully the use of state-of-the-art speaker recognition techniques for this task. Two major classification schemes in this field are *Gaussian Mixture Model – Universal Background Model* (GMM-UBM) [39] and *i-Vectors* [40] which are usually employed in combination with phonatory and articulatory information of the speaker. In this study several PD automatic detectors are analysed using GMM-UBM and *i-Vectors* in combination with different parameterizations and speech tasks.

The paper is organized as follows: Section 2 summarizes the main guidelines of this study. Section 3 develops the theoretical background about the different parameterizations and classification techniques. Section 4 introduces the experimental setup and describes the databases used in this study. Section 5 presents the obtained results. Lastly, Section 6 presents the discussions and 7, the conclusions and future work.

2. Overview and contribution

The present work performs a thorough study about the influence of the different parameters and configurations of state-of-the-art speaker recognition techniques for the detection of PD. Mainly, different combinations of acoustic material, parameterization and classification schemes are analysed separately to identify the strengths and weaknesses of each one in PD detection.

As it is depicted in Fig. 1, speech materials can be a sustained vowel, a DDK task or two different sentences. Three families of

[3] DDK tests consist in the repetition of words or syllables starting in a calm syllable rate which is increased until the speaker reaches her/his limit rate. Literature reports significant differences between controls and PD patients in several measures over DDK tasks [22]. DDK is of great interest because this task implies alternating articulatory movements where the employment of plosive syllables and different points of articulation promotes a good scenario in which changes in velocity of articulation can facilitate PD detection or assessment.

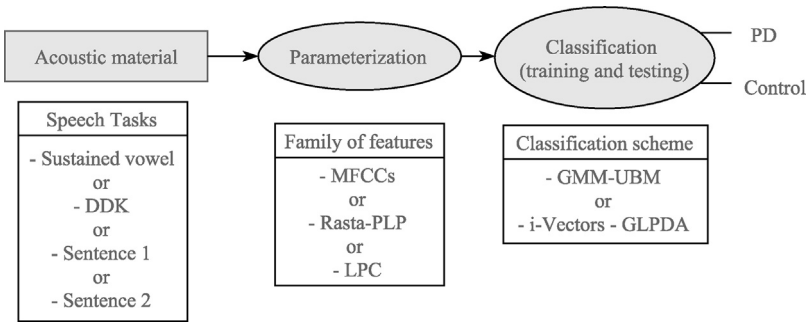

Fig. 1. Overview of methodology.

features are used separately, namely, MFCC, RASTA filtered Perceptual Linear Predictive coefficients (Rasta-PLP) and Linear Prediction Coefficients (LPC) as well as two classification schemes, GMM-UBM and *i-Vectors*. Each of these parts has some degrees of freedom such as window length, number of feature coefficients or number of Gaussians among others. This study proposes a large amount of combinations of these degrees of freedom, each one generating a different model which is tested and assessed in terms of accuracy for the detection of PD. In this way, it is possible to compare the performance of the classification methods and the suitability of the parameterizations and the speech tasks for the PD dysarthria detection. However, this study considers experiments using the different combinations separately not only to assess their convenience in PD dysarthria detection but to establish a baseline for comparison in future works.

It is important to remark that although literature shows some examples of the use of these techniques for the detection of PD [24,41,42], none of the works provides a deep study of them assessing thoughtfully the diverse degrees of freedom.

Additionally, a preliminary study of the kinetic changes is performed, since the speed and acceleration of the features which model the vocal tract can provide relevant information about articulation and its perturbations. In all of the cited combinations, this kinetic information is added to the feature vectors and the longitude of the speech segments employed to calculate them is evaluated in order to analyse its influence in the detection of PD. Particular attention has been paid to this degree of freedom which has not been thoroughly contemplated in the literature, independently of the features or classification scheme used, and on spite of the fact that fluctuations of articulation's velocity and acceleration are firmly associated to PD.

In short, the contribution of this work is twofold, firstly to explore the possibilities of state-of-the-art speaker recognition techniques in the detection of PD establishing a baseline for future works. And secondly to assess the influence of kinetic changes in this detection.

3. Theoretical background

In this section, the main techniques and basis used in the methodology, i.e. the feature families, kinetic changes and classification schemes, are introduced along with a critical discussion of its use in automatic PD detection.

3.1. Parameterizations

A great number of short term parameterizations are employed in speaker recognition tasks but literature shows that the most recurrent features in this area are LPC, MFCC, and Rasta-PLP. These three techniques are selected for this study since all of them have been widely used for speaker and speech recognition tasks and have demonstrated their capabilities to characterize the vocal tract during speech production. Each one provides a different characterization of the speech, allowing to extract some conclusions after the comparison of the different performances in the PD automatic detection with the proposed classification methods.

The first ones, LPC, were developed in the seventies [43] and are typically used to compress the audio signal, reducing the bit rate in communications. This type of feature extraction portrays largely the vocal tract and the radiation at mouth through the use of a time-variant all-poles filter which can describe the spectral envelope. LPC coefficients have not been tested in PD automatic detection. Nevertheless, LPC do not capture glottal source features and these can reveal important indicators about PD as it is stated in several works [34,23], thus, the use of LPC in this study is employed as an indicator evaluating the importance of glottal source variability in PD detection with the proposed methods.

MFCC are very well known in speech processing since the eighties [44] and can be considered as the standard speech parameterization approach. To obtain these coefficients, the spectral estimate of the signal is filtered in multiple bands and mapped in the mel scale. The logarithm of the resulting bands is transformed through the Discrete Cosine Transform (DCT) leading to the MFCC. Regarding voice and speech pathology detection, this parameterization has been successfully employed in a large number of studies with different purposes [45–48] including the detection of PD [38,42,24]. It is difficult to provide a direct interpretation of the information contained in these coefficients but it has been demonstrated that they can model the tract during articulation with some traces of glottal source-related features [49].

PLP coefficients were proposed in the nineties [50] as a cepstral domain representation of the vocal tract during articulation. To the authors of this study knowledge, these coefficients have never been used for PD automatic detection. PLP analysis of speech is based on a linear prediction algorithm but in this case, the all-pole modelling is applied to an auditory spectrum, which is more consistent with human hearing. This auditory spectrum is obtained after warping the original spectrum into the Bark scale and applying a post-processing composed of two main steps: firstly, a pre-emphasis is applied simulating the equal-loudness curve at 40 dB to consider the sensitivity of human hearing at different frequencies and secondly, a dynamic range compression is employed to compensate the non-linear relationship between the level of sound and its perceived loudness. This processing provides emphasis on the speech characteristics which are more relevant to the human hearing system and allows a low order in the subsequent all-pole model representing the vocal tract. Additionally, a RASTA filter can be included between the Bark filterbank and the auditory post-processing to obtain the RASTA-PLP. The idea of RASTA filtering is to eliminate all the components that do not contain phonetic information. To this end, a band-pass filtering is employed, alleviating the effect of convolutional noise introduced in the channel

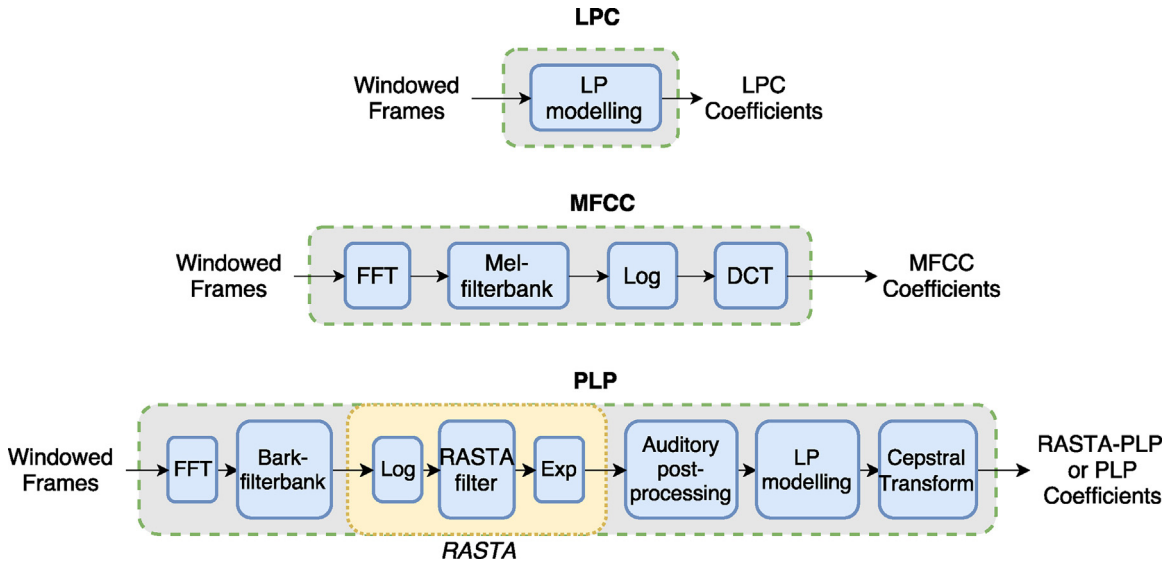

Fig. 2. LPC, MFCC and PLP calculation procedure scheme.

while smoothing some of the fast frame-to-frame spectral changes present in the short-term spectral estimate due to the analysis of artifacts [51]. The RASTA processing is a technique mainly used to mitigate the influence of noise in the codification, similar to others such cepstral mean substraction. However, while cepstral mean subtraction compares the current analysis window against the average of the whole utterance, RASTA uses a relatively short history of the signal on the order of several hundreds of milliseconds. Such a short history effectively enhances transitions between different speech segments and makes the result dependent on the previous short segment of speech such as phoneme or syllable [51].

A diagram explaining the procedure to calculate LPC, MFCC and PLP coefficients is introduced in Fig. 2. A common trait of the three feature families is that they rely on the characterization of the resonance properties of the supralaryngeal vocal tract, which has been recognized as one of the most important sources of information in speech recognition applications [52]. However, the three groups of features are defined in different domains: LPC in the frequency and PLP and MFCC in the cepstral domain. It is convenient to remark that LPC do not use perceptual mechanisms to obtain coefficients while MFCC and PLP use perceptual approximations at some point but with different approaches. Thus, including the three parameterizations in the study allows to compare the use of perceptual vs. non-perceptual and cepstral vs. non-cepstral parameterizations. In this sense, perceptual techniques can condense information while non-perceptual can provide information discarded by the other techniques. Moreover, the use of coefficients which do not include glottal source information such as LPC allows to evaluate the importance of the presence of glottal source representation in the whole process.

With independence of the family of features used, it is possible to represent a vector $\boldsymbol{p}_n$ of F features calculated in a parameterization front-end for each time window under analysis as:

$$\boldsymbol{p}_n = (p_1, \ldots, p_i, \ldots, p_F) \tag{1}$$

where n and i are the time window and coefficient indexes respectively.

3.2. Kinetic changes

Additionally to the three families of features, a characterization of the kinetic changes of each instantaneous coefficient is considered in this work in order to include variations of velocity and acceleration, which are supposed to describe deviations in articulation.

One of the first attempts to include kinetic information in a speaker recognition system is reported in [53]. There, the author describes an automatic speaker verification system using cepstrum coefficients over short segments that are expanded by an orthonormal polynomial representation aiming to add dynamical information to the system. This expanded polynomial representation is also named *delta* (Δ) and describes the slope or velocity of the instantaneous coefficients. *Deltas* are typically calculated within a certain context segment of length $\tau_{derivative}$ (derivative segment length). In particular for the referenced work, *delta* features are defined over 90 ms segments containing 9 windows of 10 ms without overlapping since that interval length is found adequate for preserving transitional information between phonemes. In a further work by the same author [54], a speech recognizer relying on cepstral characteristics and *delta* features computed through polynomial regression is introduced. Using windows of 8 ms, the performance of the system is explored within different derivative segments varying from 3 to 11 windows for the *delta* coefficients computation. Results indicate that the inclusion of the *delta* significantly improved performance compared to just considering the instantaneous coefficients. Moreover, derivative segments of 72 ms – or 9 windows – are found to be the optimal value that derives effective regression coefficients. From this point on, literature has adopted the use of these kinetic changes in a wide selection of applications due to their complementarity with the information given by instantaneous coefficients. Likewise, several variations have been proposed, such as the inclusion of double-*delta* ($\Delta\Delta$) coefficients to characterize the curvature or acceleration.

Derivatives are calculated employing an anti-symmetric FIR filter and particularly for this work, the impulse response is obtained using the first order orthogonal polynomial as described in [54]:

$$D_{i,n} = \frac{\sum_{m=-n_0}^{n_0} C_{i,n}(m) \cdot m}{\sum_{m=-n_0}^{n_0} m^2} \tag{2}$$

where $D_{i,n}$ is the derivative at the nth window, $C_{i,n}(m)$ is a function defined in the interval $-n_0 \leq m \leq n_0$ containing $2 \cdot n_0 + 1$ coefficients centred at window n and relative to the *ith* coefficient of a certain feature family (for instance, the second MFCC). Thus, the number of windows used in the FIR filter and therefore the length of the signal segment employed in this process depends on n_0.

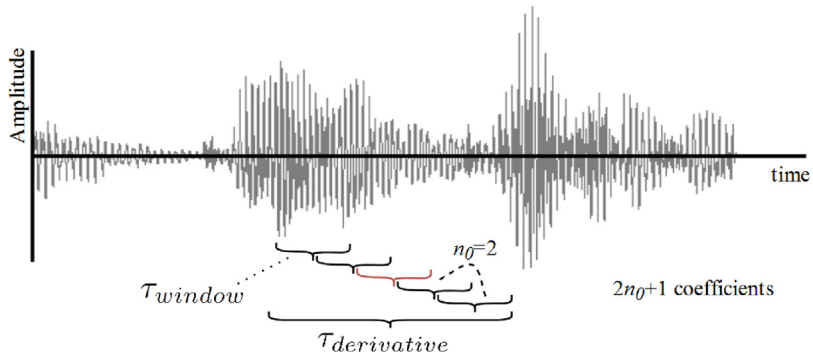

Fig. 3. Scheme illustrating $\tau_{derivative}$ and its relationship with τ_{window} and the number of coefficients used to calculate Δ for $n_0 = 2$.

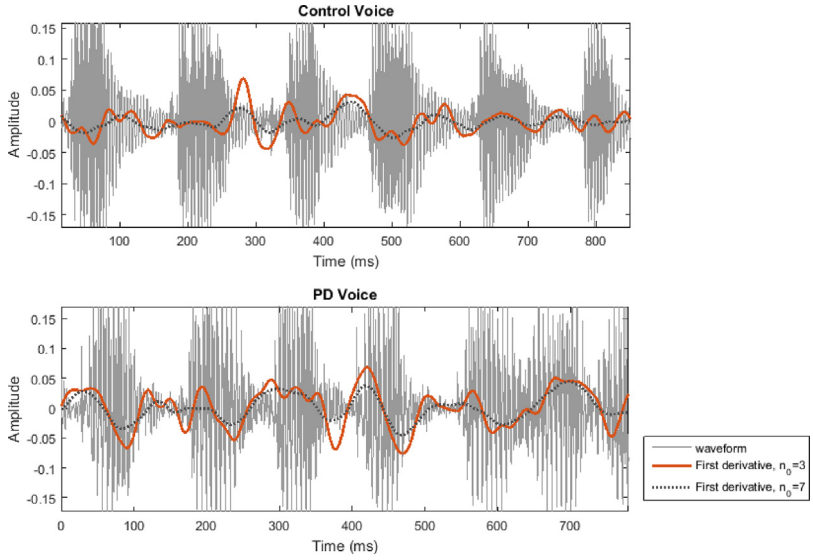

Fig. 4. First derivative of the fourth PLP coefficient ($F = 12$) calculated using two different values of n_0 during two /pa-ta-ka/ utterances of a control (upper) and a parkinsonian voice (down).

By varying this value it is possible to select the more appropriate derivative segment length or $\tau_{derivative}$ for specific purposes. Short values of $\tau_{derivative}$ might result in a noisy derivative sequence and too long values could smooth it, leading to loss of information. Fig. 3 illustrates the relationship between $\tau_{derivative}$ and the number of coefficients and τ_{window} used to calculate Δ. Fig. 4 serves as an example of the use of different n_0 values in the derivative calculation where the higher n_0, the smoother the derivative curve. Traditionally, works using derivatives do not explore the optimum n_0 and the number of FIR coefficients is not mentioned or is fixed to a value which is independent of the window size. As $\tau_{derivative}$ is dependent of n_0 and τ_{window}, using fixed n_0 when utilizing a range of different τ_{window} produces the variation of $\tau_{derivative}$ according to the changes of the window length. In other words, the use of short τ_{window} generates shorter derivative segments than those obtained when using long τ_{window}. This methodology could be unorthodox when trying to compare the performance of different τ_{window} as these two parameters – τ_{window} and $\tau_{derivative}$ – should be studied independently.

The first derivative – Δ, velocity – is obtained applying the expression (2) to the features vector and the second derivative – $\Delta\Delta$, acceleration – by using the same expression employing the first derivative array to extract $C_n(m)$. Therefore, after the parameterization of a frame, a new feature vector including derivatives can be represented as:

$$\boldsymbol{x}_n = \{\boldsymbol{p}_n, \Delta, \Delta\Delta\} = (x_1, \ldots, x_i, \ldots, x_D) \tag{3}$$

where $D = 3 \cdot F$. Then, for an utterance composed of N windows, the cluster of feature vectors, $\boldsymbol{X}_u$ is:

$$\boldsymbol{X}_u = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \ldots, \boldsymbol{x}_N\} \tag{4}$$

3.3. *Classification schemes: GMM-UBM and* i-Vectors

In the present study two classification schemes, GMM-UBM and *i-Vectors*, are used. GMM-UBM have been employed in the last few years in speaker recognition and other speech-based tasks [39], having had great popularity until the appearance of the *i-Vectors* approach [40]. The *i-Vectors* technique has demonstrated to be successful in speaker recognition and in other speech-related tasks where some examples are diarization [55], language identification [56] or accent recognition [57]. The fundamentals of these techniques are explained next.

The general idea of a GMM model is to estimate the probability density function that characterizes the cluster of vectors of dimension D of a certain class, c, using a linear combination of G multivariate Gaussian components. The model can be used to calculate the likelihood of any new vector belonging to this class. This model is represented by $\boldsymbol{\Theta}^{\boldsymbol{c}} = \{\lambda_g^c, \boldsymbol{\mu}_g^c, \boldsymbol{\Sigma}_g^c\}|_{g=1}^G$, where λ_g^c are the mixture weights $0 \leq \lambda_g^c \leq 1$, $\sum_{g=1}^G \lambda_g^c = 1$; $\boldsymbol{\mu}_g^c$ are D-dimensional means and $\boldsymbol{\Sigma}_g^c$, $D \times D$- dimensional covariance matrices.

Thus, the likelihood that one feature vector $\boldsymbol{x}_n$ belongs to class c, given the model $\boldsymbol{\Theta^c}$ is:

$$p(\boldsymbol{x}_n|\boldsymbol{\Theta^c}) = \sum_{g=1}^{G} \lambda_g^c \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_g^c, \boldsymbol{\Sigma}_g^c) \quad (5)$$

where $\mathcal{N}(\cdot)$ represents Gaussian density functions.

The model $\boldsymbol{\Theta^c}$, with c indicating membership to a certain *class* $c : c \in \{control, PD\}$, is typically estimated using a certain amount of speech windows belonging to class c following the maximum likelihood procedure using the iterative expectation-maximization algorithm [58]. When a large training set is supplied, this procedure provides a reliable estimation of the parameters of the model. However, the quality of the estimation is compromised when the training set is limited. It is possible to partially circumvent this by employing a larger auxiliary database which includes most of the speech characteristics under study, and which is modelled using a GMM procedure. This auxiliary model, $\boldsymbol{\Theta_{ubm}} = \{\lambda_g^{ubm}, \boldsymbol{\mu}_g^{ubm}, \boldsymbol{\Sigma}_g^{ubm}\}|_{g=1}^{G}$, is referred as *universal background model* (UBM) [39] and serves as a well-trained initialization, for which it is possible to adapt specific models using the provided – and more scarce – training data. The derived models, $\boldsymbol{\Theta^c_{ubm}} = \{\lambda_g^{ubm}, \hat{\boldsymbol{\mu}}_g^c, \boldsymbol{\Sigma}_g^{ubm}\}|_{g=1}^{G}$, are termed GMM-UBM and are expected to behave better than those developed using the training data directly. The adaptation from $\boldsymbol{\Theta_{ubm}}$ to $\boldsymbol{\Theta^c_{ubm}}$ is simplified adapting only the mean, $\boldsymbol{\mu}_g^{ubm}$ to $\hat{\boldsymbol{\mu}}_g^c$, using a maximum a posteriori algorithm (MAP) [39] and setting the covariance matrix and the weights to the values of the UBM, as it is typically done in speech related applications.

Therefore, the likelihood that the feature vector $\boldsymbol{x}_n$ belongs to class c, given the model $\boldsymbol{\Theta^c_{ubm}}$ is obtained using expression (5) replacing $\boldsymbol{\Theta^c}$ with $\boldsymbol{\Theta^c_{ubm}}$ and their respectives means and covariances.

The final decision about the membership of an input utterance to a certain class is taken by establishing a threshold over the score assigned to it, which is calculated by means of expression (6). The threshold is typically set at the equal error rate (EER) point determined using the training data.

$$\Lambda_{GMM-UBM} = \frac{1}{N} \log \prod_{n=1}^{N} \frac{p(\boldsymbol{x}_n|\boldsymbol{\Theta^{control}_{ubm}})}{p(\boldsymbol{x}_n|\boldsymbol{\Theta^{PD}_{ubm}})} \quad (6)$$

Another scheme to map the speakers into the two classes under study can be achieved rearranging the identity vector or *i-Vectors* approach [40]. The scheme employs the idea of GMM-UBM but using one model, $\boldsymbol{\Theta^u_{ubm}} = \{\lambda_g^{ubm}, \hat{\boldsymbol{\mu}}_g^u, \boldsymbol{\Sigma}_g^{ubm}\}|_{g=1}^{G}$, for each utterance instead for each class. Then, a transformation maps the means $\hat{\boldsymbol{\mu}}_g^u$ to a $(G \cdot D)$-dimensional vector $\boldsymbol{m}$ called *supervector* stacking mean vectors as follows:

$$\boldsymbol{m} = \{\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_G\}$$

For a utterance $\boldsymbol{X}_u$, the class- and utterance-specific-supervector $\boldsymbol{m}_u^c$, might be rearranged as:

$$\boldsymbol{m}_u^c = \boldsymbol{m}^{ubm} + \boldsymbol{T}\boldsymbol{w}_u^c \quad (7)$$

where $\boldsymbol{T}$ is a rectangular matrix of low rank called *total variability matrix*; $\boldsymbol{w}_u^c$ is a random vector of dimension q (also called *i-Vector*) having a standard normal prior distribution $\mathcal{N}(0, 1)$; and $\boldsymbol{m}^{ubm}$ is the UBM mean supervector.

Since the *i-Vectors* scheme only models data in the low dimensional total variability space, compensation techniques are often applied to address the presence of variability factors affecting performance. One popular technique is the *Gaussian Probability Linear Discriminant Analysis* (GPLDA), which further decomposes the class and utterance dependent *i-Vectors* as [59]:

$$\boldsymbol{w}_u^c = \overline{\boldsymbol{w}} + \Phi \boldsymbol{y}^c + \epsilon_u \quad (8)$$

where Φ is an eigenvoice matrix, $\overline{\boldsymbol{w}}$ is the mean *i-Vector* class- and utterance-dependent, $\boldsymbol{y}^c$ is a vector of latent factors of dimension h and the residual ϵ_u contains the channel component and describes the within-speaker variability.

To compute the class membership of a certain test *i-Vector*, $\boldsymbol{w}_{\tilde{\boldsymbol{x}}}$, with a class c, a verification trial is performed to contrast the alternative hyphotesis of the test *i-Vector* and the model having the same latent variables $\boldsymbol{y}^c$ ($\mathcal{H}_{control}$); or not ($\mathcal{H}_{PD}$). The model of the class c, $\overline{\boldsymbol{w}}^c$, is calculated as the mean *i-Vector* across the recordings belonging to c. In this manner, the score of the verification trial is as follows:

$$\Lambda_{i\text{-}Vectors} = \log \frac{p(\overline{\boldsymbol{w}}^c, \boldsymbol{w}_{\tilde{\boldsymbol{x}}}|\mathcal{H}_{control})}{p(\overline{\boldsymbol{w}}^c|\mathcal{H}_{PD})p(\boldsymbol{w}_{\tilde{\boldsymbol{x}}}|\mathcal{H}_{PD})} \quad (9)$$

These scores are calculated following the procedure proposed in [59]. The final decision is taken comparing the score with a threshold using the EER point established during the training procedure.

4. Experimental setup

This section describes the databases and the methodology employed in this work.

4.1. Acoustic material

Two databases are utilized in the present paper: the *Albayzin* and the GITA databases.

The Albayzin database [60] is a phonetically balanced dataset, sampled at 16 kHz and quantized with 16 bits, composed by a large amount of utterances in Spanish language. For this paper purposes, only the first subset from the five provided in the corpus is used to create the UBM.

On the other hand, GITA is a Colombian database presented in [61] which contains a variety of speaking conditions from 50 patients with PD and 50 control speakers, age- and sex-matched. The UPRDS, H&Y and years since diagnosis distributions of the PD patients is portrayed in Fig. 5.

In the present paper four types of speech tasks from the GITA database are used, as detailed in Table 1. Most of the tests are performed for each one of these four groups of recordings separately. The tasks comprises the DDK tests including repetitions of the syllables /pa-ta-ka/, two read sentences and a sustained vowel /a:/. The two read sentences listed in Table 1 have been selected from all the available resources since they contain a greater variety of articulation points. Tables 2 and 3 contain their phonetic transcription using the International Phonetic Alphabet (IPA) and the place of articulation of the consonants (bilabial, labiodental, dental, interdental, alveolar, palatal or velar as illustrated in [62]). These places can be anterior (bilabial, labiodental, dental, interdental and alveolar), medium (palatal) or posterior (velar). In a similar manner, both tables indicate the presence of fast and slow glides (diphthong and hiatus respectively), and 2 or 3 joint consonants. Lastly, the sustained vowel has been edited and does not include onset and offset segments. This vowel is used for some limited tests in order to compare its performance against the other speech tasks.

The GITA database is sampled at 44.1 kHz and quantized with 16 bits. Nevertheless, all recordings are filtered and downsampled to 16 kHz to match the sampling rate of the Albayzin database. This PD database is employed to train and test the different models generated through adaptation from the UBM obtained using Albayzin.

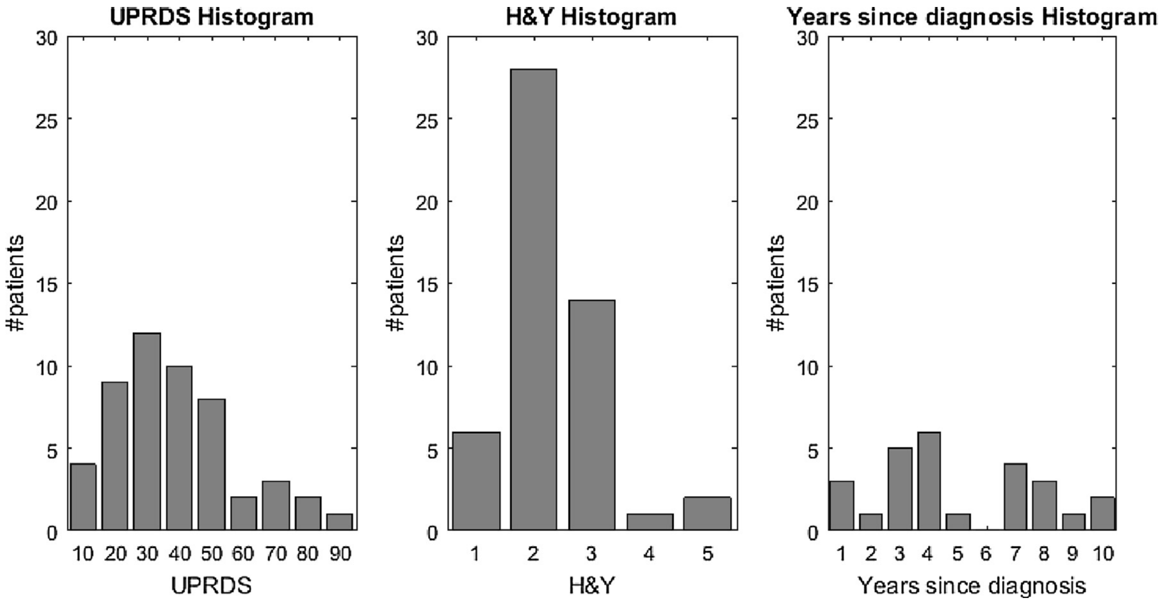

Fig. 5. UPRDS, H&Y and Years since diagnosis histograms for the PD patients within the GITA database.

Table 1
Speech tasks and phonetic transcription (IPA) from GITA database.

Name	Speech tasks and phonetic transcription
DDK	*pa-ta-ka* [pa ’ta ka]
Sentence 1	*Luisa Rey compra el colchón duro que tanto le gusta* [’lwisa’rei’kompra el kol’tʃon ’ð̞uro ke ’tan̪to le ’ɣ̞usta]
Sentence 2	*Los libros nuevos no caben en la mesa de la oficina* [los̺ ’liβ̞ros̺ ’nweβ̞os̺ no ’kaβ̞en en la ’mes̺a ð̞e la ofi’s̺ina]
Sustained vowel	*a* [a]

Table 2
Places of articulation, and number of joint vowels and consonants in Sentence 1.

Sentence 1: *Luisa rey compra el colchón duro que tanto le gusta*																																								
Phonetic transcription (IPA)	’l	w	i	s	a	’r	e	i	’k	o	m	p	r	a	e	l	k	o	l	’tʃ	o	n’	ð̞	u	r	o	k	e	’t	a	n̪	t	o	l	e’	ɣ̞	u	s	t	a
Place of articulation	Bilabial											•	•																											
	Labiodental																																							
	Dental																							•						•			•							•
	Interdental																																							
	Alveolar	•			•		•							•			•			•			•			•						•			•				•	
	Palatal																				•																			
	Velar									•								•										•									•			
Dipthong			–	–				–	–																															
Hiatus															–	–																								
2 consonants																	–	–		–	–		–	–								–	–						–	–
3 consonants												–	–	–																										

4.2. Methodology

One of the main objectives of this study is to test the state-of-the-art speaker recognition technologies in the detection of PD through speech. Nevertheless, the observation of the influence of the parameters involved in the detection, i.e. τ_{window}, F, etc., can lead to additional conclusions. For instance, the analysis of the number of FIR coefficients in the derivative calculation can point out towards finding the optimal $\tau_{derivative}$ in PD detection. In general lines, and as it is depicted in Fig. 1, a speech database is used as the input of a PD detector composed by a parameterization front-end and a classification system. These two main parts have several degrees of freedom which are varied at each training-testing cycle using independently different speech tasks. It is important to remark that thousands of specific models or detectors are built and analysed depending on the employed acoustic material, classification technique, family of features, τ_{window}, $\tau_{derivative}$, G and other degrees of freedom specific of the classi-

Table 3
Places of articulation, and number of joint vowels and consonants in Sentence 2.

Sentence 2: *Los libros nuevos no caben en la mesa de la oficina*																																										
Phonetic transcription (IPA)		l	o	ʂ	'l	i	β̞	r	o	ʂ	'n	w	e	β̞	o	ʂ	n	o	'k	a	β̞	e	n	e	n	l	a	'm	e	ʂ	a	ð̞	e	l	a	o	f	i'	ʂ	i	n	a
Place of articulation	Bilabial						•							•							•							•														
	Labiodental																																				•					
	Dental																															•										
	Interdental																																									
	Alveolar	•		•	•			•		•	•					•	•						•		•	•				•				•					•		•	
	Palatal																																									
	Velar																		•																							
Dipthong												–	–																													
Hiatus																																			–	–						
2 consonants				–	–		–	–		–	–					–	–								–	–																
3 consonants																																										

```
\MAIN PSEUDOCODE STAGE 1
for each SpeechTask do:          // SpeechTask can be: DDK, Sentence 1, Sentence 2 or Sust. vowel*
  for  τwindow = 10 ms to 40 ms do:
     for n0 = 1 to 7 ** do:
        for each FeatureFamily do:        // FeatureFamily can be: MFCC, Rasta-PLP or LPC
           for F = 10 to 20 do:      // F  = #features
            [TrainFEATURES, TestFEATURES] = CalculateFeatures {
            SpeechTask, τwindow, n0, FeatureFamily, F }
              for G = 4 to 256 do:     // G  = #gaussians
                 model = train_GMM_UBM_Model { TrainFEATURES,
                   Albayzin_DB, G }
                 results = test_GMM_UBM_Model { TestFEATURES,  model }
                 return { results }

\ENDMAIN PSEUDOCODE STAGE 1
```

Fig. 6. Pseudocode of Methodology. First stage. *Sustained vowel is used only in this stage. **n_0 does not always range between 1 and 7; in some cases these values are restricted to limit $\tau_{derivative}$ between 35 and 80 ms.

fication technique. The joint analysis of these separated models allow to evaluate the influence of the degrees of freedom in the PD automatic detection proposed in this work.

The degrees of freedom of the parameterization front-end are: family of features, F, τ_{window} and n_0. Before parameterization, all audio files from GITA database are filtered and resampled to 16 kHz. Signals from both databases are normalized and windowed using Hamming windows and τ_{window} is varied from 10 to 40 ms at 5 ms steps. Family of features are MFCC, Rasta-PLP and LPC and the number of coefficients, F, varies in the range: $\{10, \ldots, 20\}$ with steps of 2. After parameterization, derivatives are calculated using n_0 ranging from 1 to 7 avoiding $\tau_{derivative}$ values under 35 ms and over 80 ms. In that way, the use of a range of n_0 allows to compare the influence of derivatives at different τ_{window} maintaining the derivative segment lengths.

Respecting LPC and PLP, it is convenient to mention that among all the available techniques for the linear prediction algorithm, the autocorrelation method [43] is the one used in this study.

There are two main stages of methodology in which most of the different degrees of freedom or parameters are varied, depending on the classification system employed. In the first one, only GMM-UBM classification schemes are used with G taking values from 4 to 256, in powers of 2. The pseudocode of this stage can be found in Fig. 6 showing the different degrees of freedom involved, each one included in a `for` loop.

Albayzin database is used to generate the UBM, and each of the four speech tasks selected from the GITA database is employed separately for training and testing the GMM models following a k-folds cross-validation scheme, with $k = 11$. An outline of this part of the methodology is shown in Fig. 8, including the ranges of all parameters. Among all the obtained results, those with better accuracy are considered the optimal although other indicators as sensitivity, specificity, confidence interval (CI), Area Under the ROC Curve (AUC) and DET curve are also calculated in the assessment phase. CI is calculated in all cases as detailed in [63].

After this first round of modelling using GMM-UBM techniques, a second stage of training-testing with *i-Vectors*-GPLDA as classification system is performed. In this case, τ_{window} and $\tau_{derivative}$ are fixed to the values providing best results in the GMM-UBM tests as it is detailed on the pseudocode of Fig. 7 while the rest of the parameters from the parameterization front-end are varied in the same ranges as in the first stage (speech task excluding sustained vowel, family of features and F). In this scenario, the degrees of freedom of the classification system are the length of *i-Vectors* – q – which takes up values $\{30, 50, 80\}$, dimension h of the GPLDA subsystem taking values $\{2, 6, 10, 14, 18\}$ and G varying in powers of 2 from 4 to 256. An outline of this part of the methodology is shown in Fig. 9. Equally, Albayzin database is used to create the UBM. An 11-folds cross-validation procedure is performed.

In short, for the first stage of the methodology (GMM-UBM classification scheme), the degrees of freedom are speech task, family of features, F, G, window size and n_0 ($\tau_{derivative}$). Considering the best results obtained in this stage, the last two values are fixed to the optimum and are used in the second stage, (*i-Vectors* classification scheme) in which the degrees of freedom are the speech tasks (DDK and the two sentences), family of features, F, h, q and G.

The sustained vowel /a:/ is only used in GMM-UBM tests for comparison purposes. This acoustic material has poor articulation

```
\MAIN PSEUDOCODE STAGE 2
for each SpeechTask do:                    // SpeechTask can be: DDK, Sentence 1 or Sentence 2.
  for each FeatureFamily do:               // FeatureFamily can be: MFCC, Rasta-PLP or LPC.
    for F = 10 to 20 do:                   // F = #features
      [TrainFEATURES, TestFEATURES] = CalculateFeatures { SpeechTask,
      Optimal t_window, Optimal n_0, FeatureFamily, F }
      for G = 4 to 256 do:                 // G = #gaussians
        for q = 30 to 80 do:
          for h = 2 to 18 do:
            model = train_IVECTORS_GPLDA_Model { TrainFEATURES,
            Albayzin_DB,  G, q, h }
            results = testGMMModel { TestFEATURES, model }
            return { results }
\ENDMAIN PSEUDOCODE STAGE 2
```

Fig. 7. Pseudocode of methodology. Second stage.

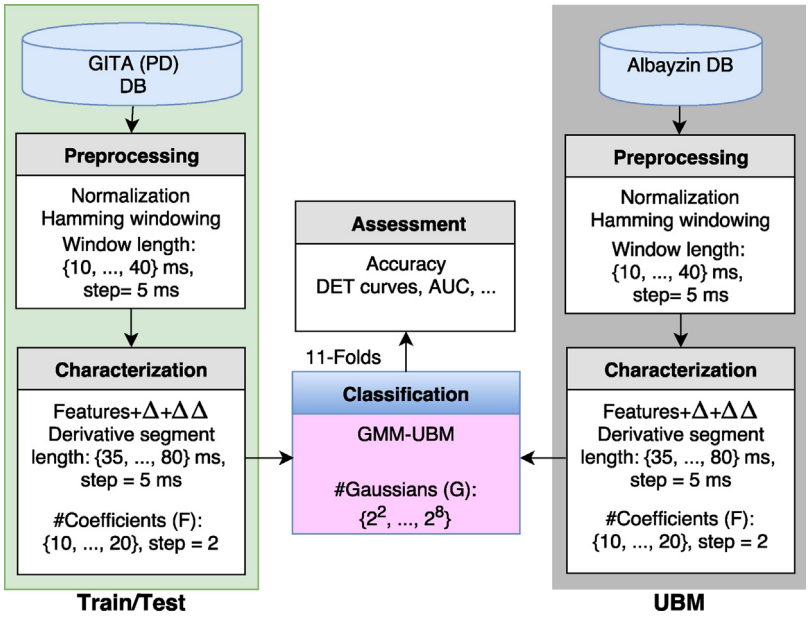

Fig. 8. GMM-UBM training methodology outline. Features can be MFCC, Rasta-PLP and LPC and derivative coefficients Δ + ΔΔ.

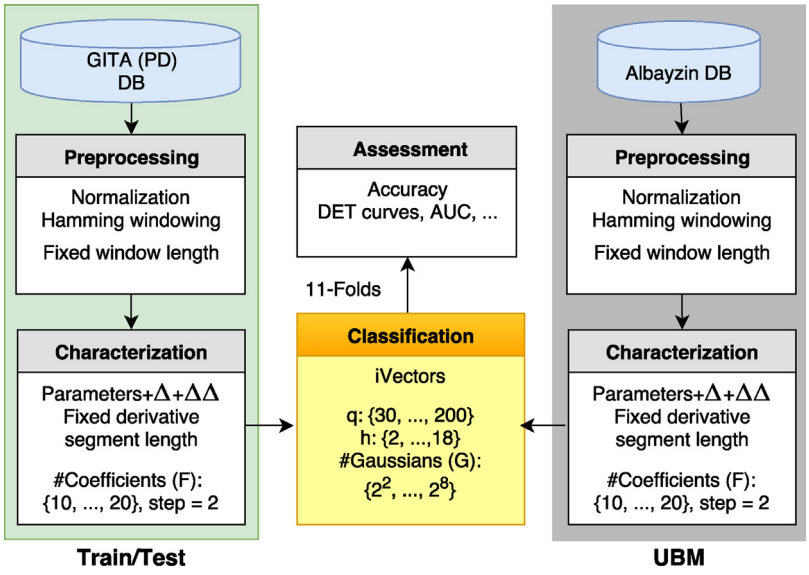

Fig. 9. *i-Vectors* training methodology outline. Features can be MFCC, Rasta-PLP and LPC with derivative coefficients Δ + ΔΔ.

Table 4
Number of variations of each degree of freedom (Speech task, τ_{window}, n_0, F, G, q and h) in the two classification schemes and the three feature families.

Number of variations for each degree of freedom				
Classification scheme	Degrees of freedom	Family of features		
		LPC	MFCC	Rasta-PLP
GMM-UBM (Stage 1)	Speech task	4	4	4
	τ_{window}	7	7	7
	n_0	1–5	1–5	1–5
	# Coefficients (F)	6	6	6
	# Gaussians (G)	7	7	7
i-Vectors GPLDA (Stage 2)	Speech task	3	3	3
	τ_{window}	1	1	1
	n_0	1	1	1
	# Coefficients (F)	6	6	6
	# Gaussians (G)	7	7	7
	q	3	3	3
	h	5	5	5

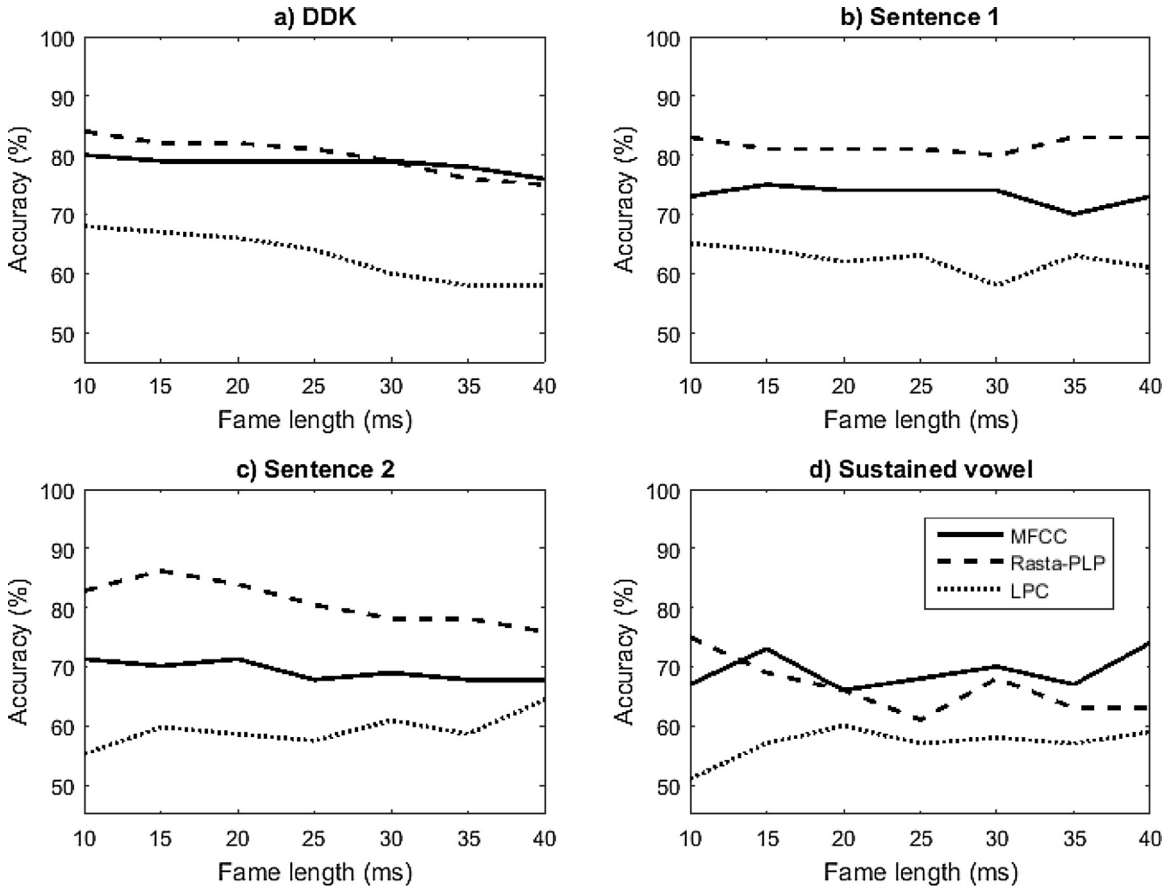

Fig. 10. Best accuracies in the range {10, . . ., 40} ms of τ_{window} for the three different parameterizations, GMM-UBM classification techniques and using speech tasks: (a) DDK, (b) Sentence 1, (c) Sentence 2, and (d) Sustained vowel. These values are obtained using different number of G, F and $\tau_{derivative}$ at each point.

information which, in principle, is an important aspect for PD detection (as detailed in Section 1). Moreover, its use in combination with the *i-Vectors* scheme would not lead to more conclusions beyond those obtained in the GMM-UBM scheme. Although some works have proven that sustained vowels can carry information about the presence of PD, these utilize the onset and offset segments of the vowels during speech production and specific methodologies different to those included in this study which are out of the scope of this work.

In both stages, all the feature vectors of one specific speaker (X_u) are included in only one fold among all of the generated to perform the cross-validation. Coefficients, x_i, from the training folds are normalized in the interval [−1, 1] while the weights obtained in this normalization process are applied to the test fold.

Table 4 shows the number of possible variations of each degree of freedom in the two stages, observing the two classification schemes and, for each one, the three parameterization families. For instance, as q can take values {30, 50, 80}, the number of variations is 3. The number of variations of n_0 range between 1 and 5 as the values that it can take depend on τ_{window}. Each combination and fold leads to a new model. In the GMM-UBM scheme, an amount of 2856 models are trained and tested for each family of features and fold, while 1890 models are obtained in the *i-Vectors* stage. There-

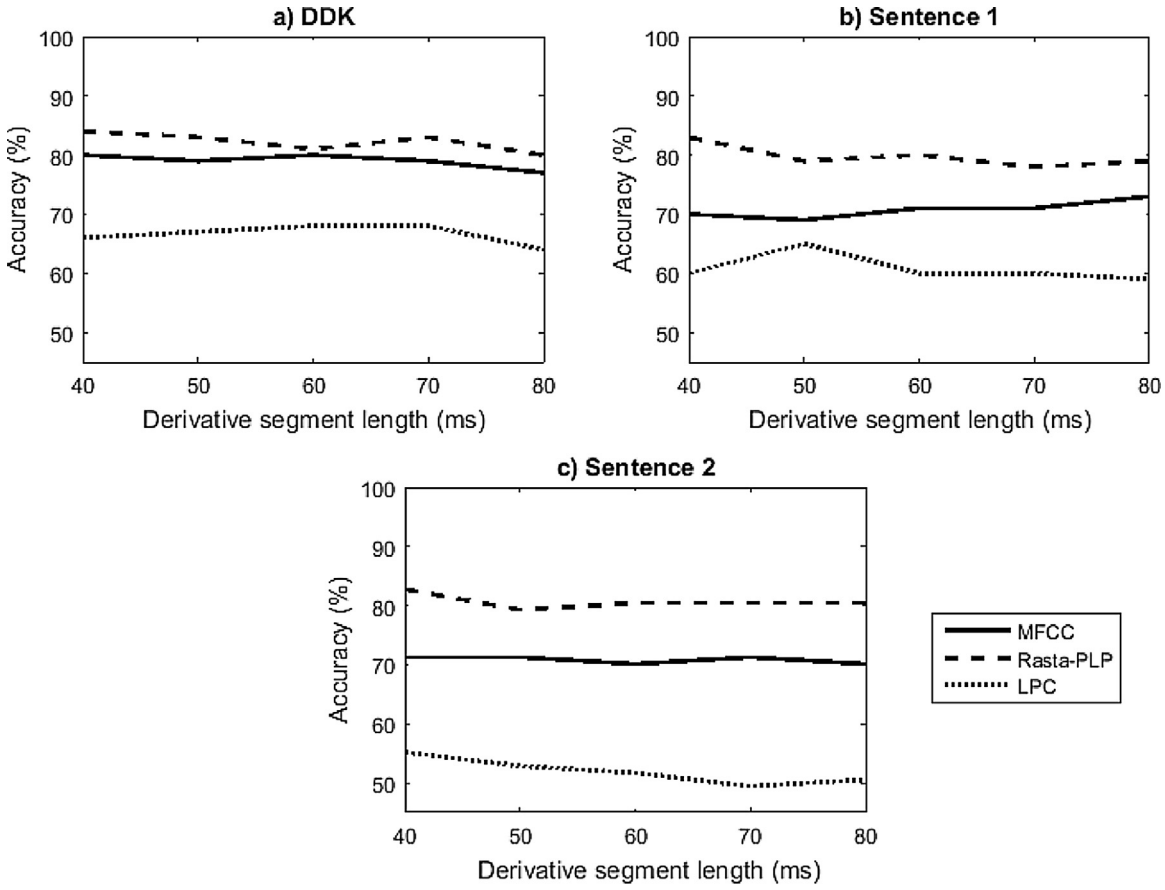

Fig. 11. Best accuracies using τ_{window} = 10 ms as a function of $\tau_{derivative}$ for the three different parameterizations using speech tasks: (a) DDK, (b) Sentence 1, and (c) Sentence 2. These values are obtained using different number of *G* and *F* at each point.

fore, the total number of models trained and tested in the whole study is 156618, 14238 models per fold.

Additionally, and in order to assess the influence of kinetic changes in global results, new tests are made using only the configurations leading to the best results in the two stages (family of features, *G*, *F*, etc.) but excluding derivatives from the feature vector. In the same manner, mutual information [64] of all features and kinetic changes respect to control and PD classes is calculated.

5. Results

For the sake of simplicity in the presentation, this section only includes results leading to best accuracy values and those showing a possible influence of the parameters in performance justified by the presence of PD.

5.1. *Influence of* τ_{window}, n_0 *and* F *in the first stage of methodology*

In order to analyse the influence of τ_{window} in the accuracy of the system, Fig. 10 shows the best accuracy results as a function of τ_{window} and family of features using a GMM-UBM classification scheme.

In the same manner, to evaluate the impact of n_0 variation on accuracy with respect to the speech task materials and parameterizations, Fig. 11 shows the best results in terms of accuracy as a function of n_0 when using τ_{window} = 10 ms, which is the optimum value in most of the cases as can be inferred from Fig. 10. In order to evaluate the influence of the number of coefficients of the FIR filter when using different window lengths, Fig. 12 depicts accuracy as a function of $\tau_{derivative}$ considering only Rasta-PLP + Δ + ΔΔ and employing 10, 15 and 20 ms window lengths.

Additionaly, Fig. 13 shows the maximum accuracy obtained in the first stage as a function of the number of coefficients, considering τ_{window} = 10 ms.

5.2. *GMM-UBM and* i-Vectors *global results*

Best absolute accuracy in the first stage (86%) is obtained using τ_{window} = 15 ms and n_0 = 2. Therefore, the two parameters are fixed to these optimum values for the tests in the second stage (*i-Vectors* scheme).

The best global accuracy results in the two stages of methodology are included in Table 5. This table allows to compare the best outcomes depending on the speech task, parameterization and classification technique. Focusing on the sentence 2, which leads to the best absolute accuracy, DET curves for the three parameterizations and the two classification techniques are calculated (Fig. 14). Fig. 15 shows accuracy results as a function of *G* and *F* with the two different classification techniques and Rasta-PLP, which is the family of features which leads to best outcomes.

Finally, Table 6 shows the best global accuracy obtained for each speech task including their AUC, specificity and sensitivity and the specific configuration leading to these results. The values of AUC, specificity and sensitivity are defined within the interval [0, 1]. Best global results are obtained using Sentence 2, Rasta-PLP + Δ + ΔΔ, *i-Vectors*, q = 50 and h = 14. It is possible to observe that Accuracy, AUC, Specificity and Sensitivity are the highest of all tests, having Specificity and Sensitivity similar values. Fig. 16 shows DET curves for those results referred in the same table.

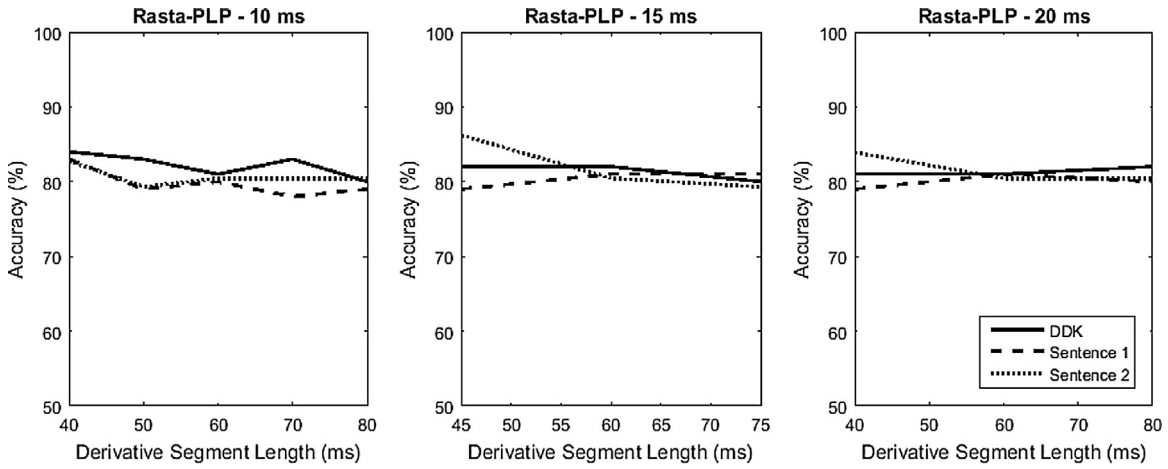

Fig. 12. Best accuracies using $\tau_{window} = \{10, 15, 20\}$ ms as a function of $\tau_{derivative}$ for DDK, Sentence 1 and Sentence 2 using Rasta-PLP + Δ + ΔΔ coefficients. These values are obtained using different number of *G* and *F* at each point.

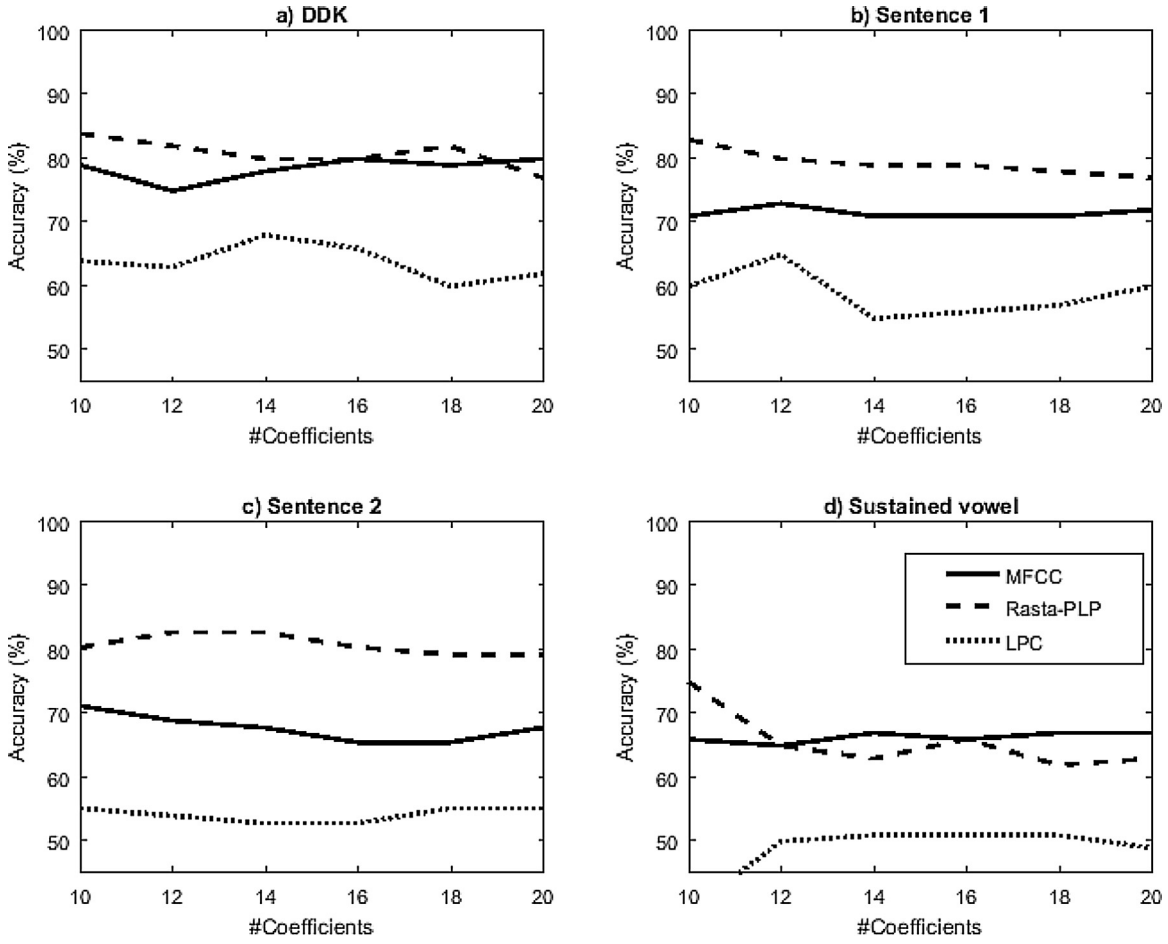

Fig. 13. Best accuracies using $\tau_{window} = 10$ ms as a function of the number of coefficients for the three different parameterizations, GMM-UBM classification techniques and using speech tasks: (a) DDK, (b) Sentence 1, (c) Sentence 2, and (d) Sustained vowel. These values are obtained using different number of *G* and $\tau_{derivative}$ at each point.

Table 5
Best accuracy results ± CI for the different acoustic materials using MFCC, Rasta-PLP and LPC with derivative coefficients Δ + ΔΔ. The highest accuracies are marked in bold.

Best accuracy results ± CI (%)						
Speech task	MFCC + Δ + ΔΔ		Rasta-PLP + Δ + ΔΔ		LPC + Δ + ΔΔ	
	GMM-UBM	*i-Vectors*	GMM-UBM	*i-Vectors*	GMM-UBM	*i-Vectors*
DDK	80 ± 8	80 ± 8	84 ± 7	82 ± 8	68 ± 9	74 ± 9
Sentence 1	77 ± 8	75 ± 8	83 ± 8	79 ± 8	65 ± 9	79 ± 8
Sentence 2	71 ± 10	77 ± 9	86 ± 7	**87 ± 7**	64 ± 10	71 ± 10
Sustained vowel /a:/	74 ± 9	–	75 ± 8	–	60 ± 10	–

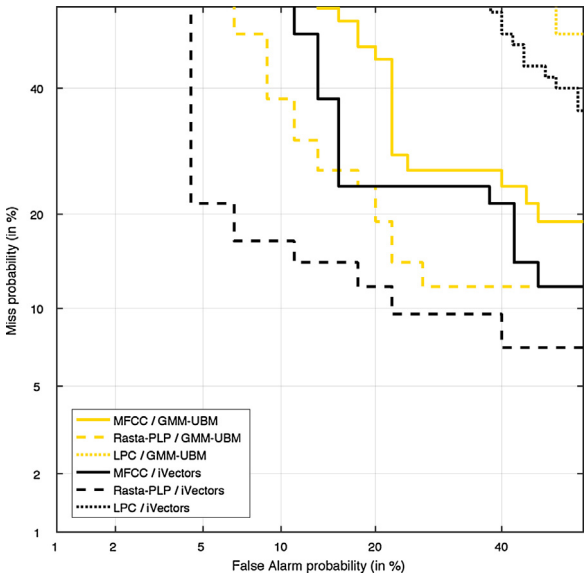

Fig. 14. DET curves for MFCC, Rasta-PLP and LPC parameterizations + Δ + ΔΔ leading to the highest accuracies using the two classification techniques in Sentence 2 speech task tests.

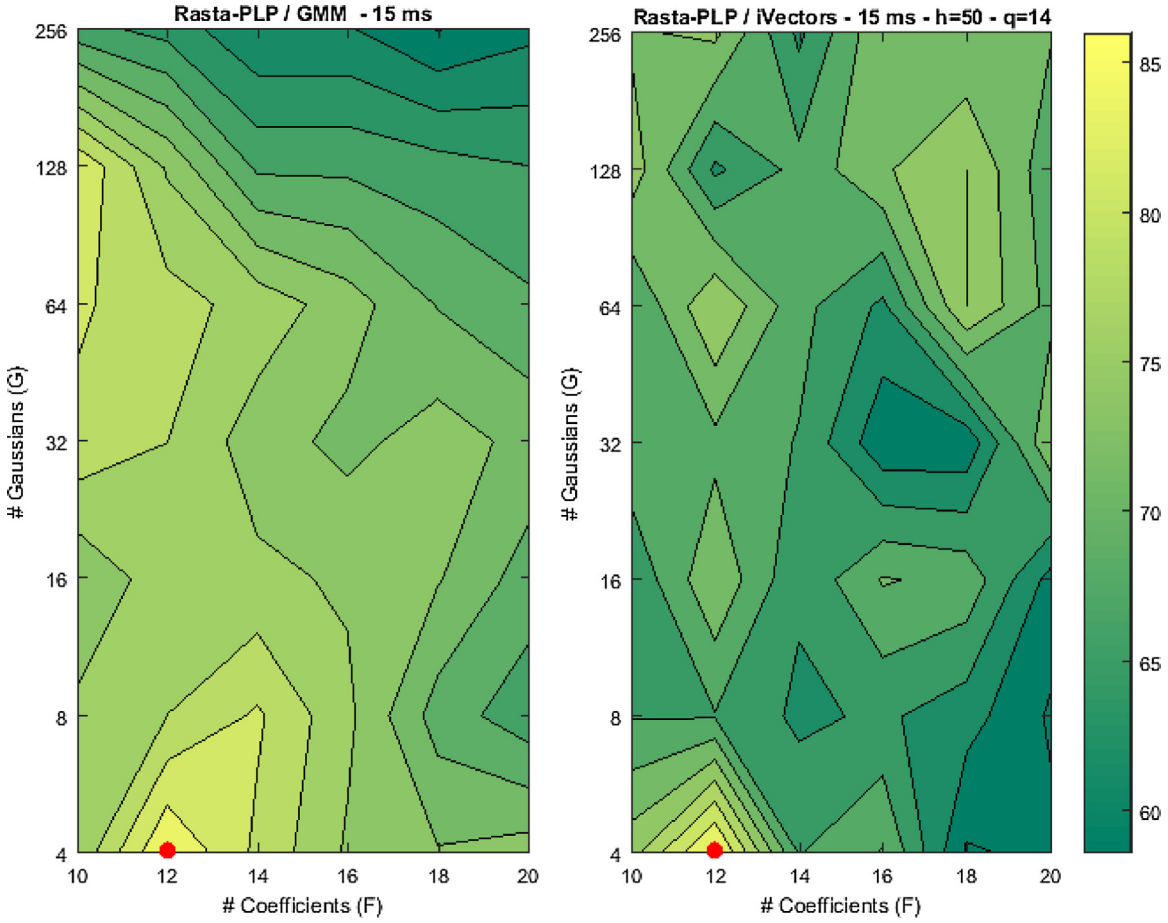

Fig. 15. Best accuracies in GMM and *i-Vectors* tests using τ_{window}=15, n_0 = 2 with Rasta-PLP + Δ + ΔΔ for Sentence 2 as a function of *G* and *F*. In the case of *i-Vectors*, *h* = 50 and *q* = 14. Best points of operation are marked with a red dot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Additionally, Fig. 17 shows mutual information of features and derivatives leading to best results (Rasta-PLP and *i-Vectors* as referenced in Table 6). In the same way, Table 7 includes the results of this configuration with and without derivatives in the feature vector.

For the sake of simplicity, the influence of *q* and *h* in the results is not reported in detail since it has been considered that any conclusions regarding the performance of these parameters would be barely linked with the presence of PD. It is not possible to obtain any solid conclusions concerning *q* and *h* as the values providing

Table 6
Best results for each acoustic material. The configuration leading to the highest accuracy is marked in bold.

Best results										
Speech task	Accu. ± CI (%)	AUC	Spec.	Sens.	Feature family	τ_{window} (ms)	n_0	Class. technique	# Gauss.	# Coefs.
DDK	84 ± 7	0.89	0.84	0.84	Rasta-PLP + Δ + ΔΔ	10	3	GMM-UBM	8	10
Sentence 1	83 ± 7	0.86	0.86	0.80	Rasta-PLP + Δ + ΔΔ	10	3	GMM-UBM	64	10
Sentence 2	**87 ± 7**	**0.93**	**0.89**	**0.86**	**Rasta-PLP + Δ + ΔΔ**	**15**	**2**	***i-Vectors***	**4**	**12**
Sust. vowel	75 ± 8	0.83	0.70	0.80	Rasta-PLP + Δ + ΔΔ	10	3	GMM-UBM	256	10

Table 7
Results using Rasta-PLP and the configuration leading to best results with and without derivatives to analyse its influence in accuracy, AUC, specificity and sensitivity.

Feature family	Classification technique	Accuracy (%)	CI (%)	AUC	Specif.	Sensit.
Rasta-PLP + Δ + ΔΔ	GMM-UBM	86	7	0.86	0.87	0.93
Rasta-PLP (without derivatives)	GMM-UBM	71	9	0.68	0.74	0.78
Rasta-PLP + Δ + ΔΔ	*i-Vectors*	87	7	0.93	0.89	0.86
Rasta-PLP (without derivatives)	*i-Vectors*	71	9	0.74	0.68	0.79

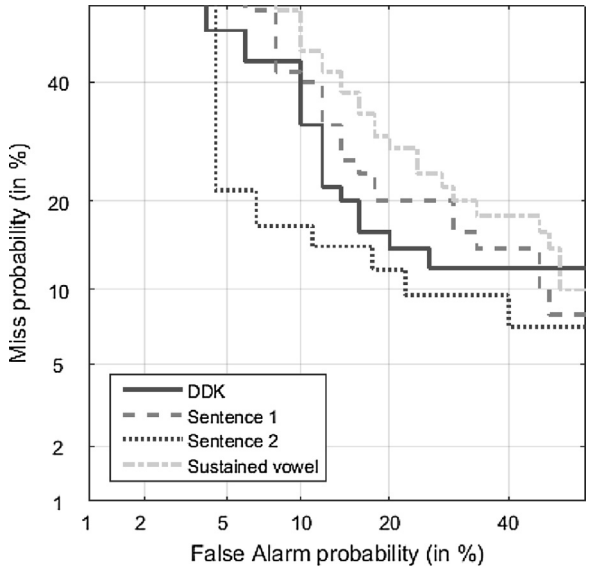

Fig. 16. DET curves for the different speech tasks leading to the highest accuracies and the configurations detailed in Table 6.

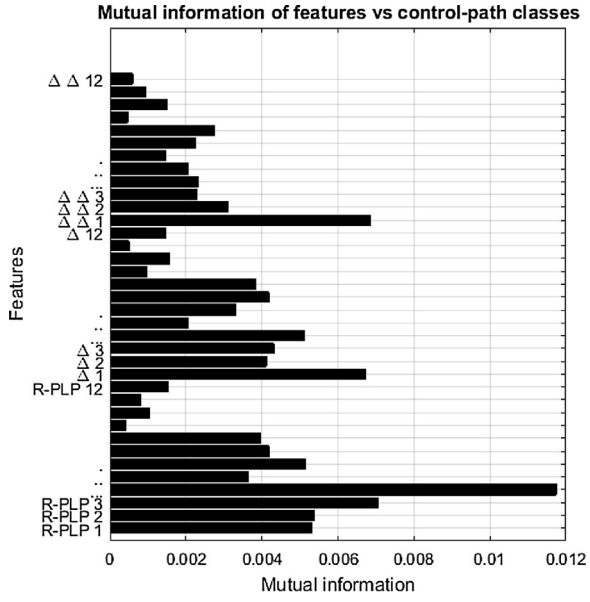

Fig. 17. Mutual information of Rasta-PLP (R-PLP) features with respect to the control/pathological classes.

best results are more influenced by the length of the database (i.e. number of speakers) and the length of the feature vector utilized rather than by the particularities of parkinsonian speech.

6. Discussion

The exposed results allow to analyse the influence of parameterization, τ_{window}, kinetic changes, speech task and classification schemes for the automatic PD detection through speech. The employed methodologies are the state-of-the-art techniques for speaker recognition but different optimum configuration parameters are expected for PD detection.

Influence of the family of features. As it can be inferred from Tables 5 and 6, best results are obtained using Rasta-PLP + Δ + ΔΔ coefficients, independently of the speech task and the classification techniques employed. In same cases, the performance of the MFCC family is similar to the one achieved with Rasta-PLP but providing a lower maximum accuracy. On the other hand, the differences in performance between LPC and the other two families of features are remarkable. Figs. 10, 11 and 13 show that best results for LPC are usually lower than those obtained with Rasta-PLP and MFCC and in many cases are at 20 absolute points under the best Rasta-PLP accuracies.

Going into detail to the Rasta-PLP parameterization, to evaluate the influence of RASTA filtering in the whole process, some prospective tests are performed using PLP with and without RASTA filtering to compare its influence in PD detection. To this aim, the GMM-UBM methodology depicted in Fig. 8 is repeated limiting $n_0 = 4$ using DDK speech tasks and a maximum τ_{window} of 30 ms. Fig. 18 shows the best accuracy vs. window length obtained in the prospective tests to analyse the use of RASTA filter, while the best numeric results are included in Table 8. As it can be inferred from these outcomes, RASTA filtering does not provide remarkable improvements to PLP suggesting that this filtering does not enhance the capabilities of PLP for the detection of PD in this study.

Influence of τ_{window} and $\tau_{derivative}$. With reference to τ_{window}, in general, best results are obtained with the shortest windows, 10 and 15 ms, which is consistent with the adoption of quasi-stationary conditions. Something similar occurs with the derivative segment length in which a high $\tau_{derivative}$ could blur the velocity and acceleration tracking. Fig. 11 shows the accuracy obtained with the three parameterizations as a function of $\tau_{derivative}$ for $\tau_{window} = 10$ ms. In this scenario, best results are obtained using $n_0 = 3$, which corresponds with the use of 7 coefficients in the FIR filter and a $\tau_{derivative}$ of 40 ms. Using different τ_{window}, the optimum $\tau_{derivative}$ is similar, as it can be inferred from Fig. 12 in which the derivative segment lengths providing best results are most of

Table 8
Results using PLP feature family with and without RASTA filter in order to its influence in accuracy, AUC, specificity and sensitivity.

Parameterization	τ_{window} (ms)	Accuracy ± CI (%)	AUC	Specif.	Sensit.
PLP + Δ + ΔΔ	15	81 ± 8	0.87	0.82	0.80
Rasta-PLP + Δ + ΔΔ	10	82 ± 8	0.87	0.82	0.82

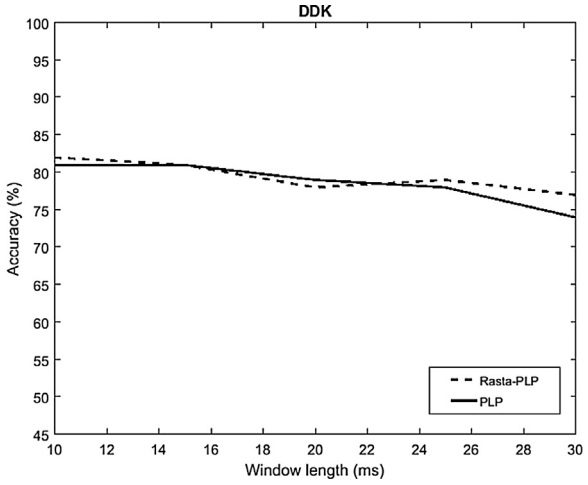

Fig. 18. Best accuracies varying τ_{window} in the range 10, . . ., 30 ms for PLP parameterization with and without RASTA filtering, using DDK speech task and GMM-UBM classification technique.

the times around 40 ms. This suggests that the analysis of velocity changes in parkinsonian speech characterization should be performed within these ranges. This length allows more resolution to characterize imprecise articulation coordination than larger segments, which is coherent with the dysarthria features found in [23,65].

Influence of kinetic changes. A supplementary analysis of the relevance of kinetic changes in the group of features is performed. As derivatives characterize the velocity and acceleration of vocal tract movements, the analysis of its relevance can help disclosing the importance of articulation in the studied detectors. Firstly, Fig. 17 shows mutual information of all the coefficients leading to the best results (as detailed in Table 6, Sentence 2) and the speaker's condition. The mutual information relative to kinetic coefficients is significant with regard to the mutual information of static Rasta-PLP coefficients. Besides, several tests are repeated introducing some variations to assess the influence of these derivatives in the accuracy. These new tests are performed employing 12 Rasta-PLP coefficients without derivatives, τ_{window} = 15 ms, and the two classification schemes using values of {4, 8, 16, 32} for *G*. In the case of *i-Vectors*, *q* and *h* are varied in the same range specified in the methodology of the study. After removing derivatives, tests give a maximum of 71% of accuracy in both cases as shown in Table 7, confirming the influence of the kinetic changes and the importance of articulation.

Influence of speech tasks. Regarding the speech tasks and taking into account the overlapping margins in accuracy caused by the CI, the results suggest that Sentence 2 leads to better accuracies, specially when parameterizing the signal following the Rasta-PLP technique. Observing Tables 2 and 3, it is possible to infer that the amount of changes between velar, alveolar and labial articulation places is not significantly different than those found in Sentence 1. Moreover, the bilabial-dental-velar articulation changes which occur recurrently in the DDK task do not provide better outcomes than those obtained with Sentence 2, even when the former task requires a higher excursion of the articulation organs which is repeated again and again. Nevertheless, one of the most noteworthy difference between Sentence 2 and the rest of tasks is that it includes more fricative consonants, produced by means of the narrowing of the articulatory organs which do not touch each other. However, as the differences in accuracy between the three running speech tasks are not dramatic it is not possible to completely conclude that Sentence 2 is more reliable for future tests. In any case, it is not clear if this sentence includes some segments or allophones more relevant in the detection of PD, which should be studied in further works.

In that sense, the models trained with the sustained vowel are the least efficient among those obtained in the study. These results support the importance of articulation for PD automatic detection with speaker recognition techniques. As it is introduced in Section 1, the works centred in phonatory aspects providing good results make use of the phonation of different vowels or extract vowel segments from running speech, which contain certain articulatory information. In the present case none of these circumstances occur, which explains why accuracy does not exceed 75% with the sustained vowel task as shown in Fig. 10d). However, works like [23,34] suggest that phonation by itself can provide information about tremor or noise attributable to PD, but to take advantage of these circumstances, the use of features characterizing glottal source is necessary. Going into detail to Fig. 10d), results with LPC show poor accuracy as expected since this family of features does not contain relevant information about the glottal source. Better outcomes are achieved using MFCC and Rasta-PLP coefficients which suggest that both families can characterize in some manner the glottal source, and that could justify the relatively good results that they provide (75% accuracy) using only one sustained vowel.

Influence of classification techniques. Regarding the classification techniques, GMM-UBM methods give generally the best results with all acoustic materials but Sentence 2, which provides the best outcomes combined with *i-Vectors*. In general, the differences of best accuracy results comparing both schemes are small. Moreover, although the *i-Vectors*-GPLDA methodology is more sophisticated and usually produces better performance for speaker recognition, it is more oriented to use larger UBM and training databases, containing more variability than the ones used in this work. The fact that the best results provided by this last technique employing relatively short *i-Vectors* ($q = 50$) while typical values in speaker recognition are around $q = 400$ is influenced by the small size of the database and by the low inter-class variability, caused by PD. Moreover, it is possible to observe in Fig. 15 that the highest accuracy results are more spread in the GMM case (left) with respect to the *i-Vectors* best results (right), which are concentrated into a small region indicating that the former technique could generalize better.

Particularities of the patients in the GITA database. The histograms on Fig. 19 are referred to the errors in the PD class using the configuration leading to the best results (87% of accuracy). These show the percentages of false rejection in PD class relatives to the total number of false rejections with respect to UPRDS, H&Y and years since diagnosis. It can be inferred from the histograms that most of the errors occur normally in the early and mid stages. The detection in advanced stages is more efficient although there are some errors for UPRDS equal to 60 (considering that there are only two patients in this gap) or in patients with more than 7 years since diagnosis. In any case, the results suggest that there are common factors or perturbations in the voice and speech of early and

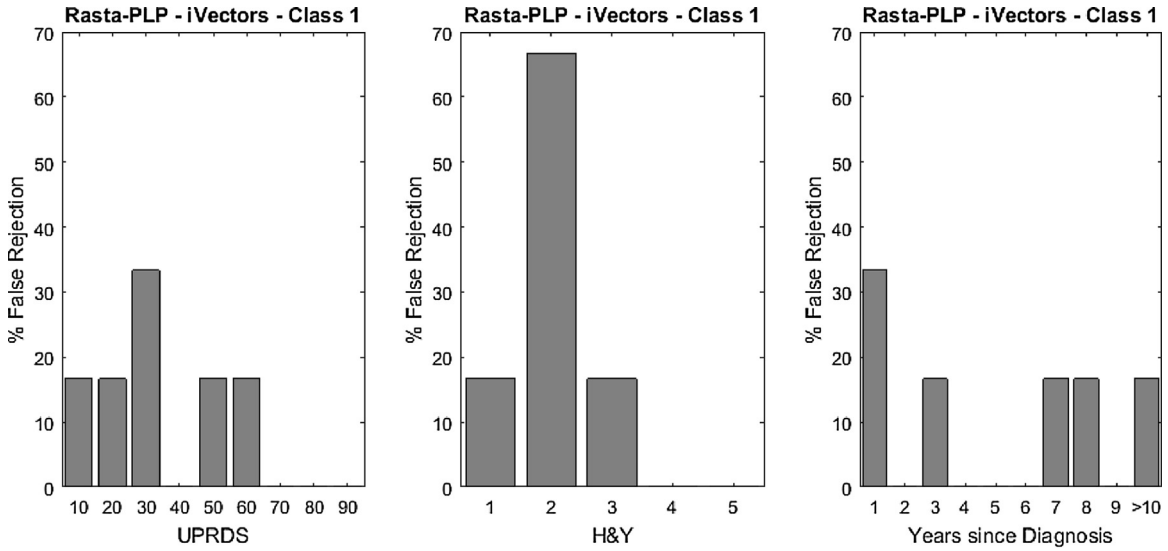

Fig. 19. % False rejection distribution in PD class relative to total false rejection of PD class respect to UPRDS, H&Y and years since diagnosis in the best-results case, using Rasta-PLP and *i-Vectors* in Sentence 2 as detailed in Table 6.

advanced stages, as the same system is able to detect most of the cases of both.

Although the minimum error found in all the analysed detectors is 13%, the theoretical limit of false rejection in a speech-based PD detector is not clearly delimited. Some specialists suggest that a 90% of PD patients suffer from dysarthria after a median latency period of 7 years since diagnosis [66,12]. These considerations would limit the false rejection in automatic detectors to 10%. Recent works like [67] have studied the presence of signs in the voice of PD patients and have quantified the percentage of affected patients to 100% using the Robertson dysarthria profile [68] in a cohort of 48 patients in several stages. However, these works are focused in perceptual estimations or preliminary quality of voice analysis. None of them have studied in detail the perturbations in voice during the early stages neither the differences between the dysarthric disturbances caused by PD and by ageing. Thus, it is not possible to compare results with the minimum theoretical limit but this value depends on the stage in which the patients included in the database are, and it could be more than 0% as the percentage of patients without perceived dysarthria (most of them in early stages) but with deviations in articulation is unknown.

Other considerations. It is known that the three families of features provide information about the vocal tract and the added $\Delta + \Delta\Delta$ coefficients can capture these changes during articulation. Nevertheless, LPC parameterizes the vocal tract in the frequency domain, whereas PLP and MFCC provide the information in the cepstral domain and use perceptual modelling. These two facts seem to be relevant in the speech processing for the detection of PD. Additionally, the presence of glottal source information in MFCC and, possibly, in Rasta-PLP could contribute to these good results. Taking into account the curves depicted in Fig. 10d), in which results are similar in MFCC and Rasta-PLP, there could be a causality relationship between the possible existence of glottal source information in PLP (as in MFCC) and the differences in the results obtained using the sustained vowel respect to LPC, but this existence cannot be completely confirmed with the performed tests. On the other hand, the optimum configuration provides values of 87% in accuracy. This value is similar to most of those reported in the state-of-the-art of PD detection. For instance, in [23], values up to 88% of accuracy are obtained. Other works as [30] obtain values up to 80% using vowels containing articulatory information extracted from monologues, the same value reached in [38] when combining phonatory, prosodic and articulatory features for PD automatic detection. Nevertheless, results are difficult to compare when different databases are employed, taking into account that in most of the cases they differ in language, sex and age distribution, speaker tasks and number of speakers. Moreover, there are not too many works analysing automatic detectors of PD based only on voice or speech. Some works provide hybrid methods combining speech with other sources such as posture and gait [69] and others use very small and unbalanced databases [70,71]. This makes unfeasible to perform too many comparisons with the obtained results at the moment.

Also, the literature provides preliminary attempts of using automatic speaker recognition procedures such as *i-Vectors* and GMM-UBM to detect or assess PD [41,42,24]. Some of these schemes employ multiple features but without a thoughtful consideration of the degrees of freedom or a comparison of results between different methodologies. For instance, [24] uses MFCC and GMM-UBM techniques with running speech to detect PD but employs the same database for UBM modelling and training-testing, which could bias the results. Moreover, the work does not analyse the different parameterization and classification ranges or degrees of freedom. In this sense, the study of the optimum parameters for speaker recognition has been addressed in multiple works and these have been reconsidered in the automatic PD detection scenario in this work.

Regarding the CI obtained on this study, in most of the cases it is relatively high in comparison to those obtained in state-of-the-art speaker recognition works in which typical values are around 0.5%. These CIs could be enhanced in two manners. The first one could be improving the accuracy results although, as mentioned, the false rejection could have a theoretical limit. The second one can be achieved using more speakers, but obtaining large databases of speech from PD patients implies an important cost. That is why most of pathological voice (or speech) databases are usually small in comparison to speaker or speech detection aimed databases.

Lastly, it is convenient to contemplate that, as it is considered in [14], some symptoms as depression can influence the speech rate and therefore articulation. In the recording of the GITA database the presence of depression in patients is not assessed and, thus, this could bias the results.

7. Conclusions and future work

In this work, state-of-the-art speaker recognition techniques are applied and adapted to a different application domain: the detection of PD using the patient's speech. Three families of features are considered, MFCC, Rasta-PLP and LPC along with their respective derivatives, utilizing multiple configurations. Equally, two classification techniques, namely GMM-UBM and *i-Vectors*, are used to train and test automatic detectors. The objective of this study is mainly twofold: firstly to evaluate the potential to apply these techniques to a new application scenario analising their different degrees of freedom to establish a baseline to compare results with further studies; and secondly, to evaluate the influence of kinetic changes of instantaneous coefficients and the importance of the number of FIR coefficients for derivatives in the detection of PD.

While best results, 87% of accuracy, are obtained using Rasta-PLP parameterization and read sentences, none of the classification techniques seems to stand out respect to the other. However, the potential of *i-Vectors* is not exploited totally since this technique is more aimed to work with larger and more complex databases, including more variability. An adaptation of this methodology for small databases could be advisable in future works. As results suggest that the variability between the two studied classes is small, techniques to reduce intra-class variance should be applied in the future to improve classification results.

The tuning of the $\tau_{derivative}$ for velocity calculation provides a margin of gain of more than 5 absolute points in classification, concluding that the optimum value is around 40 ms. In the future, the study of kinetic changes and particularly the tuning of acceleration must be addressed and its influence in the global detector, evaluated. New features based on velocity and acceleration related to articulation should be considered.

Additionally, although the signs of PD appear on each patient in a particular manner, it could be advisable to analyse its influence on specific types of phonemes in order to identify articulation places or movements more likely to include perturbations caused by the disease. Moreover, as automatic detection of PD is achieved essentially by means of the detection of parkinsonian dysarthria, further studies should analyse the differences between the influence of this type of dysarthria and the one caused by non-parkinsonian neurological disorders.

Finally, it is possible to conclude that the state-of-the-art techniques for speaker recognition provide good accuracy results in the automatic detection of PD and can be considered as a baseline in future works. Results suggest that Rasta-PLP characterization provides the best accuracies compared to MFCC and LPC using short τ_{window} (10–15 ms) and read sentences as acoustic material. The kinetic changes have to be taken into account in the future as their relevance has been proven in the automatic detection of PD patients.

Acknowledgements

The authors of this paper want to thank to Jesús Francisco Vargas Bonilla and Julián David Arias-Londoño from the Faculty of Engineering at Universidad de Antioquia who cooperated with Juan Rafael Orozco-Arroyave in the recording of the GITA database. This work was supported by the Ministry of Economy and Competitiveness of Spain (grants EEBB-I-17-12092, BES-2013-062984 and project *TEC2012-38630-C04-01*), Universidad Politécnica de Madrid (*Ayudas para la realización del doctorado – RR01/2011*, *XV Ayudas Consejo Social* and *Ayudas EEBB para PDI*) and Ministry of Education of Spain (PRX15/00385), with special thanks to the Fulbright Foundation.

References

[1] R.F. Pfeiffer, Z.K. Wszolek, M. Ebadi, Parkinson's Disease, CRC Press, 2013.
[2] A.J. Hughes, S.E. Daniel, Y. Ben-Shlomo, A.J. Lees, The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service, Brain 125 (4) (2002) 861–870.
[3] S. Fahn, Recent Developments in Parkinson's Disease, Raven Pr, 1986.
[4] M.M. Hoehn, M.D. Yahr, Parkinsonism onset, progression, and mortality, Neurology 17 (5) (1967), 427–427.
[5] F.L. Darley, A.E. Aronson, J.R. Brown, Differential diagnostic patterns of dysarthria, J. Speech Lang. Hear. Res. 12 (2) (1969) 246.
[6] H. Ackermann, W. Ziegler, Articulatory deficits in parkinsonian dysarthria: an acoustic analysis, J. Neurol. Neurosurg. Psychiatry 54 (12) (1991) 1093–1098.
[7] J. Kegl, H. Cohen, H. Poizner, Articulatory consequences of Parkinson's disease: perspectives from two modalities, Brain Cogn. 40 (2) (1999) 355–386.
[8] P. Blanchet, G. Snyder, Speech rate deficits in individuals with Parkinson's disease: a review of the literature, J. Med. Speech – Lang. Pathol. 17 (1) (2009) 1–7.
[9] T. Tykalova, J. Rusz, J. Klempir, R. Cmejla, E. Ruzicka, Distinct patterns of imprecise consonant articulation among Parkinson's disease, progressive supranuclear palsy and multiple system atrophy, Brain Lang. 165 (2017) 1–9.
[10] J.W. Tetrud, Preclinical Parkinson's disease detection of motor and nonmotor manifestations, Neurology 41 (5 Suppl. 2) (1991) 69–71.
[11] G. Weismer, Philosophy of research in motor speech disorders, Clin. Linguist. Phon. 20 (5) (2006) 315–349.
[12] J.R. Duffy, Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, Elsevier Health Sciences, 2013.
[13] B.T. Harel, M. Cannizzaro, P.J. Snyder, Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study, Brain Cogn. 56 (1) (2004) 24–29.
[14] S. Skodda, U. Schlegel, Speech rate and rhythm in Parkinson's disease, Mov. Disord. 23 (7) (2008) 985–992.
[15] S. Skodda, W. Grö, U. Schlegel, W. Grönheit, U. Schlegel, Intonation and speech rate in parkinson's disease: general and dynamic aspects and responsiveness to levodopa admission, J. Voice 25 (4) (2011) e199–205.
[16] A.B. Walsh, Basic parameters of articulatory movements and acoustics in individuals with Parkinson's disease, Mov. Disord. 27 (7) (2012) 843–850.
[17] K. Tjaden, J. Lam, G. Wilding, Vowel acoustics in Parkinson's disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions, J. Speech Lang. Hear. Res. 56 (5) (2013) 1485–1502.
[18] J. Illes, E. Metter, W. Hanson, S. Iritani, Language production in Parkinson's disease: acoustic and linguistic considerations, Brain Lang. 33 (1) (1988) 146–160.
[19] D. Van Lancker Sidtis, K. Cameron, J.J. Sidtis, Dramatic effects of speech task on motor and linguistic planning in severely dysfluent parkinsonian speech, Clin. Linguist. Phon. 26 (8) (2012) 695–711.
[20] A. Benitez Burraco, E. Herrera, F. Cuetos, A core deficit in Parkinson's disease? Neurologia 31 (4) (2016) 223–230.
[21] D.V.L. Sidtis, J. Choi, A. Alken, J.J. Sidtis, Formulaic language in Parkinson's disease and Alzheimer's disease: complementary effects of subcortical and cortical dysfunction, J. Speech Lang. Hear. Res. 58 (5) (2015) 1493–1507.
[22] H. Ackermann, I. Hertrich, T. Hehr, Oral diadochokinesis in neurological dysarthrias, Folia Phoniatr. Logop. 47 (1) (1995) 15–23.
[23] M. Novotný, J. Rusz, R. Čmejla, E. Růžička, Automatic evaluation of articulatory disorders in Parkinson's disease, IEEE/ACM Trans. Audio, Speech and Lang. Process. (TASLP) 22 (9) (2014) 1366–1378.
[24] J.R. Orozco-Arroyave, F. Hönig, J.D. Arias-Londo no, J.F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, E. Nöth, Automatic detection of Parkinson's disease in running speech spoken in three different languages, J. Acoust. Soc. Am. 139 (1) (2016) 481.
[25] A.M. Goberman, M. Blomgren, Fundamental frequency change during offset and onset of voicing in individuals with Parkinson disease, J. Voice 22 (2) (2008) 178–191.
[26] C.E. Stepp, Relative fundamental frequency during vocal onset and offset in older speakers with and without Parkinson's disease, J. Acoust. Soc. Am. 133 (3) (2013) 1637–1643.
[27] S. Skodda, W. Grönheit, U. Schlegel, Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease, PLoS ONE 7 (2) (2012) e32132.
[28] Y.-I. Bang, K. Min, Y.H. Sohn, S.-R. Cho, Acoustic characteristics of vowel sounds in patients with Parkinson's disease, NeuroRehabilitation 32 (3) (2013) 649–654.
[29] J.A. Whitfield, A.M. Goberman, Articulatory-acoustic vowel space: application to clear speech in individuals with Parkinson's disease, J. Commun. Disord. 51 (2014) 19–28.
[30] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, E. Ruzicka, Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task, J. Acoust. Soc. Am. 134 (3) (2013) 2171–2181.
[31] V. Fraïle, H. Cohen, Temporal control of voicing in Parkinson's disease and tardive dyskinesia speech, Brain Cogn. 40 (1) (1999) 118–122.
[32] M. Asgari, I. Shafran, Predicting severity of Parkinson's disease from speech, in: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, IEEE, 2010, pp. 5201–5204.

[33] J. Rusz, R. Cmejla, H. Ruzickova, E. Ruzicka, Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease, J. Acoust. Soc. Am. 129 (1) (2011) 350–367.
[34] K.S. Perez, L.O. Ramig, M.E. Smith, C. Dromey, The Parkinson larynx: tremor and videostroboscopic findings, J. Voice 10 (4) (1996) 354–361.
[35] M.A. Little, P.E. McSharry, E.J. Hunter, J. Spielman, L.O. Ramig, et al., Suitability of dysphonia measurements for telemonitoring of Parkinson's disease, IEEE Trans. Biomed. Eng. 56 (4) (2009) 1015–1022.
[36] A. Tsanas, M. Little, Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests, IEEE Trans. Biomed. Eng. 57 (4) (2010) 884–893.
[37] A. Tsanas, M.A. Little, C. Fox, L.O. Ramig, Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease, IEEE Trans. Neural Syst. Rehabil. Eng. 22 (1) (2014) 181–190.
[38] T. Bocklet, S. Steidl, E. Nöth, S. Skodda, Automatic evaluation of Parkinson's speech-acoustic, prosodic and voice related cues, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2013) 1149–1153.
[39] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digit. Signal Process. 10 (1) (2000) 19–41.
[40] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (4) (2011) 788–798.
[41] G. An, D.G. Brizan, M. Ma, M. Morales, A.R. Syed, A. Rosenberg, Automatic recognition of unified Parkinson's disease rating from speech with acoustic, i-vector and phonotactic features, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2015, Janua, Dresden, Germany, 2015, pp. 508–512.
[42] J. Kim, M. Nasir, R. Gupta, M. Segbroeck, D. Bone, M. Black, Z.I. Skordilis, Z. Yang, P. Georgiou, S. Narayanan, Automatic estimation of Parkinson's disease severity from diverse speech tasks, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2015) 914–918.
[43] X. Huang, A. Acero, H.-W. Hon, R. Foreword By-Reddy, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall PTR, 2001.
[44] S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (4) (1980) 357–366.
[45] 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 Trans. Biomed. Eng. 53 (10) (2006) 1943–1953.
[46] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, A. Ghio, Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia), in: Proceedings of European Conference on Speech Communication and Technology (Eurospeech 2005), ISCA, Lisbon, Portugal, 2005, pp. 149–152.
[47] A. Gelzinis, A. Verikas, M. Bacauskiene, Automated speech analysis applied to laryngeal disease categorization, Comput. Methods Programs Biomed. 91 (1) (2008) 36–47.
[48] A. Belhaj, A. Bouzid, N. Ellouze, Edema and nodule pathological voice identification by SVM classifier on speech signal, Int. Rev. Comput. Softw. (IRECOS) 10 (5) (2015) 495–501.
[49] A. Mesaros, J. Astola, The Mel-frequency cepstral coefficients in the context of singer identification, in: 6th International Conference on Music Information Retrieval, ISMIR, London, UK, 2005, pp. 610–613.
[50] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, J. Acoust. Soc. Am. 87 (4) (1990) 1738–1752.
[51] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. Speech Audio Process. 2 (4) (1994) 578–589.
[52] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors, Speech Commun. 52 (1) (2010) 12–40.
[53] Cepstral analysis technique for automatic speaker verification, IEEE Trans. Acoust. Speech Signal Process. 29 (2) (1981) 254–272.
[54] S. Furui, Speaker independent isolated word recognition using dynamic features of speech spectrum, IEEE Trans. Acoust. Speech Signal Process. 34 (1) (1986) 52–59.
[55] M. Rouvier, S. Meignier, A global optimization framework for speaker diarization, Proceedings of Odyssey: The Speaker and Language Recognition Workshop (2012) 146–150.
[56] M. Li, S. Narayanan, Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification, Comput. Speech Lang. 28 (4) (2014) 940–958.
[57] H. Behravan, V. Hautamäki, T. Kinnunen, Factors affecting i-vector based foreign accent recognition: a case study in spoken Finnish, Speech Commun. 66 (2015) 118–129.
[58] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B (Methodol.) (1977) 1–38.
[59] D. Garcia-Romero, C.Y. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems., in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Florence, Italy, 2011, pp. 249–252.
[60] Albayzín speech database: design of the phonetic corpus, Eurospeech 1993. Proceedings of the 3rd European Conference on Speech Communication and Technology, vol. 1 (1993) 175–178.
[61] J. Orozco-Arroyave, J. Arias-Londo no, New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease., Proceedings on the International Conference on Language Resources and Evaluation (LREC) (2014).
[62] A. Quilis, Tratado de fonología y fonética espa nolas, Editorial Gredos, 1993.
[63] N. Saenz-Lechon, J.I. Godino-Llorente, V. Osma-Ruiz, P. Gomez-Vilda, Methodological issues in the development of automatic systems for voice pathology detection, Biomed. Signal Process. Control 1 (2) (2006) 120–128.
[64] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons, 2012.
[65] R.D. Kent, G. Weismer, J.F. Kent, H.K. Vorperian, J.R. Duffy, Acoustic studies of dysarthric speech: methods, progress, and potential, J. Commun. Disord. 32 (3) (1999) 141–186.
[66] J. Müller, G.K. Wenning, M. Verny, A. McKee, K.R. Chaudhuri, E.A. Jellinger, Progression of dysarthria and dysphagia in postmortem-confirmed Parkinsonian disorders, Arch. Neurol. 58 (2) (2001) 259.
[67] G. Defazio, M. Guerrieri, D. Liuzzi, A.F. Gigante, V. di Nicola, Assessment of voice and speech symptoms in early Parkinson's disease by the Robertson dysarthria profile, Neurol. Sci. 37 (3) (2016) 443–449.
[68] S.J. Robertson, F. Thomson, Speech therapy in Parkinson's disease: a study of the efficacy and long term effects of intensive treatment, Int. J. Lang. Commun. Disord. 19 (3) (1984) 213–224.
[69] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. Biglan, E. Dorsey, M. Little, Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study, Parkinsonism Relat. Disord. 21 (6) (2015) 650–653.
[70] A. Benba, A. Jilbab, A. Hammouch, Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machines, Int. J. Electr. Eng. Inform. 7 (2) (2015) 297–307.
[71] M. Shahbakhi, D.T. Far, E. Tahami, Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine, J. Biomed. Sci. Eng. 7 (2014) 147–156.