A forced gaussians based methodology for the differential evaluation of Parkinson's Disease by means of speech processing

Laureano Moro-Velazquez , Jorge Andres Gomez-Garcia , Juan Ignacio Godino-Llorente , Jesús Villalba , Jan Rusz , Stephanie Shattuck-Hufnagel , Najim Dehak

Keywords:
Speech processing
Machine learning
Parkinson's Disease
Phoneme
Forced alignment

A B S T R A C T

Literature evidences the existence of hypokinetic dysarthria in parkinsonian patients and, consequently, the objective characterization of the dysarthric signs associated to the articulatory aspect of speech can be used to detect Parkinson's Disease (PD) providing clinicians with new tools to support the clinical diagnosis.

However, no work has analyzed in detail the importance of the different phonemes in the automatic detection of PD from the speech.

This work proposes new approaches for this detection by using new classification schemes that allow to compare independently the different phonetic units of patients and controls employed during several speech tasks. Three different parkinsonian corpora were used allowing cross-validation and cross-corpora trials.

The results of cross-validation trials (k-folds) provided accuracies between 81% and 94%, with AUC between 0.87 and 0.97 depending on the corpus, while cross-corpora trials yielded accuracies between 66% and 76% with AUC between 0.76 and 0.87.

These results suggest that PD affects to the articulatory sequence as a whole, influencing more clearly phonetic units requiring a higher narrowing of the vocal tract. Additionally, text-dependent utterances are considered as the recommended speech task for the detection of PD in this type of schemes as these allow to compare more precisely the phonetic units of patients and controls. Lastly, this work discusses the existence of a glass ceiling in the accuracy of the systems for the automatic detection of PD using speech, concluding that this is below 95% for most of the cases.

1. Introduction

The neurodegenerative processes associated to Parkinson's Disease (PD) cause multidimensional voice and speech impairments called dysphonia [1], dysprosody [2,3] and dysarthria in patients, in particular *hypokinetic dysarthria* [4–7]. This type of dysarthria is characterized by a reduction of the articulation amplitude [8,9] and a decrease of intelligibility mostly, among other signs.

To this respect, several works in literature point out to an influence of PD in phoneme production in patients, especially in the case of some specific allophones.[1] One of the earliest studies trying to determine the articulatory deficits of parkinsonian patients from a phonetic point of view is [10] in which authors analyzed the dysfunctions of larynx, lips and tongue (back, tip and blades) in 200 idiopathic and postencephalitic untreated PD patients. In that work, two trained listeners evaluated the speech of the patients to perceptually assess their quality of voice and the misarticulation of

[1] It is important to discern between two concepts: phonemes and allophones. A phoneme is a phonetic unit which cannot be decomposed in minor units and which can distinguish one word from another. The term allophone is referred to each of the pronunciation variants of a certain phoneme, usually, as a function of this phoneme in a word or syllable and depending on the adjacent phonemes. For instance, the phoneme /g/ has the allophone 'g' the Spanish word "gusta" and 'ɣ' in "disgusta".

allophones. Results show that 45% of the patients exhibited lingual or labial (or both) abnormalities during articulation. The authors reported that the errors are mainly concentrated in the consonants requiring the greatest narrowing or closure during articulation, with more errors found in velar articulations, mostly /k/ and /g/ phonemes, which has been confirmed in works [11–14]. In the subsequent work [11] using the same speech corpus, authors reported incomplete contact of articulators for plosive stops and partial constrictions for fricatives. One of the main conclusions found in this publication is that the intra-speaker consistency of the errors is near to 98%, meaning that when a patient misarticulated a phoneme, this error was repeated all along the session. However, this work did not include inter-session recordings and, thus, the longitudinal intra-speaker error consistency could not be assessed. Another important observation found in [11] is that inter-speaker misarticulation consistency reached 97%, meaning that the vast majority of the patients produced the same error substitution when a phoneme was misarticulated. In most of the cases, this error consisted in the substitution of stop phonemes by fricatives, in a phenomenon known as spirantization.

In this sense, authors of a more recent work, [3], found perceptual and objective deterioration of phonatory, articulatory and prosodic aspects of speech in 80 patients, independently of the stage of the disease, indicating that abnormalities on speech can occur in early stages. These results are in concordance of those from [15] in which authors characterized articulatory deficits to detect PD in patients from early stages.

In the studies [16,17], authors performed an analysis of the influence of PD in vowels [16] and consonants [17] of the speech from 25 sentences in habitual, clear, loud and slow conditions. The first work [16], using formant-related measurements such as Vowel Space Area (VSA), reports significant differences between the patients and controls, in concordance with the findings of other studies [3,18–22]. One of the conclusions of [16] as well as [18] is that the production of the vowel /u/ during articulation is more affected by PD than /a/ or /i/. The subsequent study [17], using acoustic features characterizing the articulatory constriction during consonants only found subtle differences between the same consonants produced by the parkisnonian and control groups. This is contradictory with the findings exposed in [10–14] where a strong influence of the disease is found in certain consonants.

Despite all of the exposed evidences, no work performs a detailed study of the importance of the different phonemes in the automatic detection of PD from the speech, to the best of authors' knowledge. Consequently, this is one of the main objectives of the present work.

To this respect, Figs. 1 and 2 illustrate two examples of the possible differences in articulation and phonation between patients and controls. In the first case, Fig. 1 shows the waveform and spectrogram of the syllable /pe/ within the word "petaca" in Spanish from a newly diagnosed patient (a) with Unified Parkinson's Disease Rating Scale (UPDRS) 9, and a control speaker (b). In the case of the patient, the vowel presents more irregularities, which is reflected in the spectrogram as an absence of well defined formants. In Fig. 2, both speakers pronounce the word "es" within the Spanish sentence "La petaca blanca es mía". In the case of the patient (a) it is observed that the articulation of the phoneme /s/ becomes almost a continuation of the phoneme /e/ as it is reflected in the waveform and in the spectrogram where the /e/ formants do not disappear when /s/ is pronounced. This contrasts with the analogous waveform and spectrogram of the control speaker (b) where the two allophones are clearly separated. This *voicing leakage* effect may be caused due to an incoordination in the use of the glottal source which leads to continuous vibration of the vocal folds even during the articulation of the /s/ consonant, where an interruption of the phonation was expected.

On the other hand, in the last decade multiple works are proposing new schemes that can be used as tools to support the diagnosis of PD [1,24–26] helping clinicians to perform an earlier diagnosis, which usually can take several years [27]. The present study can be framed in this type of works.

Thus, in view of these evidences this study considers the particularities of articulatory movements in parkinsonian speakers to distinguish between patients and controls in a binary detection system. In the proposed methodology, the different phonemes were used separately to create classifiers able to discriminate between the two types of speakers, namely patients and controls, based on statistical models of the acoustic features of these phonemes. Thus, the purpose of this work is twofold: to propose new approaches to detect PD from the speech and to analyze the importance of the different phonetic units within the proposed schemes.

With these objectives, speech recognition technologies such as forced alignment [28] were used in the present work in order to segment the speech signal, obtaining the separate allophones. These techniques have been used previously in other works such as [29] to assess the intelligibility in pathological speech, example which could be considered as a precedent of this work in the use of such techniques. Then, the obtained allophones were used to train separately different Gaussians, obtaining forced-Gaussian Mixture Models (fGMM) for PD detection.

The document is structured as follows: Section 2 describes the theoretical background, introducing the concepts and techniques used in the experimental setup, described in Section 3. Section 4 presents the results and Sections 5 and 6, the discussion and conclusions respectively.

2. Theoretical background

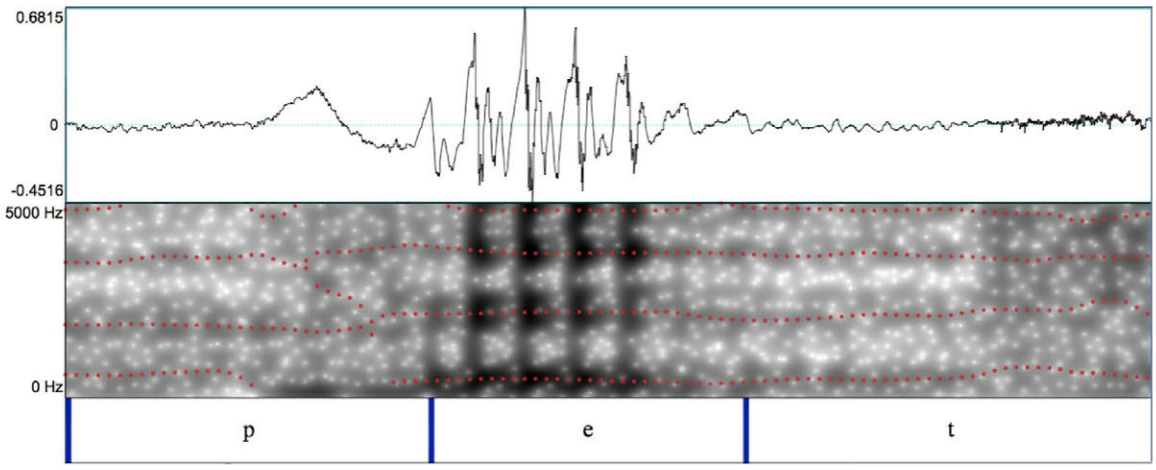
2.1. Speech forced alignment

Speech forced alignment techniques are used to automatically detect allophones in an utterance knowing its transcription (and therefore, the sequence of phonemes of the speech signal). The outcome of the forced alignment is the automatic segmentation of the signal into separated phonetic units (allophones or phonemes, depending on the system) and the labeling of each of these segments with their corresponding phonetic label. This technique can be considered as supervised since in contrast to, for example, text independent automatic speech recognition techniques, the system knows the transcription of the analyzed utterance beforehand. An example of forced alignment obtained with a model trained using Kaldi toolkit [30] is shown in Fig. 3 where it is possible to see the alignment of the sentence "La petaca blanca".

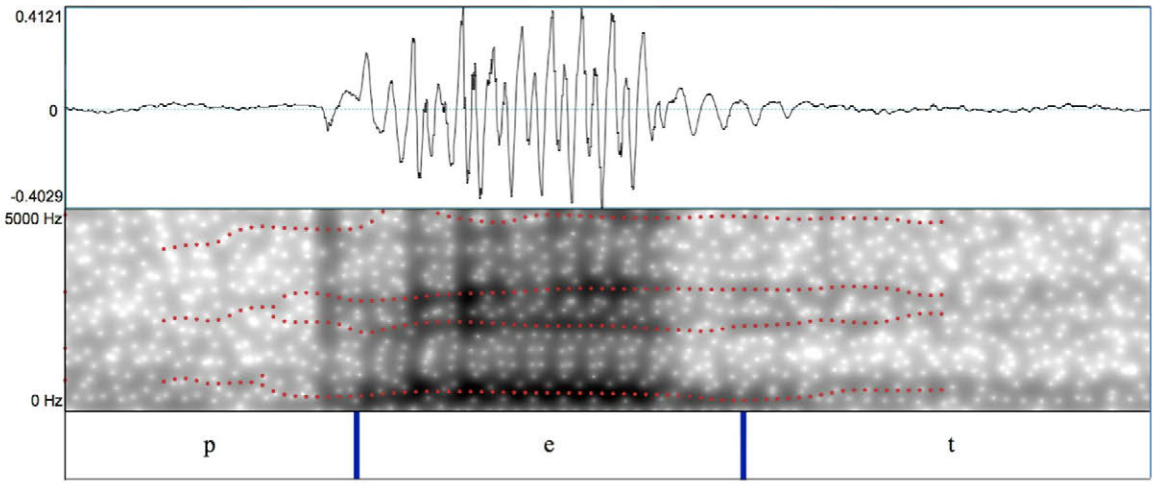
For the purposes of this work, a Forced Alignment Model (FAM) was obtained with Kaldi following a Gaussian Mixture Model (GMM) – Hidden Markov Model (HMM) architecture. In general lines, the training of this type of FAM follows these steps:

1. All the speech signals are characterized, commonly using Mel Frequency Cepstral Coefficients (MFCC) and its associated first and second derivatives ($+\Delta + \Delta\Delta$) [31] or other similar features.
2. Employing these features, a GMM-based monophone model is trained, taking into account all the possible allophones contained in the training corpus.[2] This first model consists in a group of GMM, representing each of the singular allophones or monophones. It is obtained after several iterations of model training

[2] Notice that the corpus must include the transcription files of all the utterances, and metadata such as lexicon, dictionaries and other auxiliary files.



(a) Idiopathic PD female speaker. Age: 59. UPDRS: 9. Span: 169 ms



(b) Control female speaker. Age: 59. Span: 169 ms

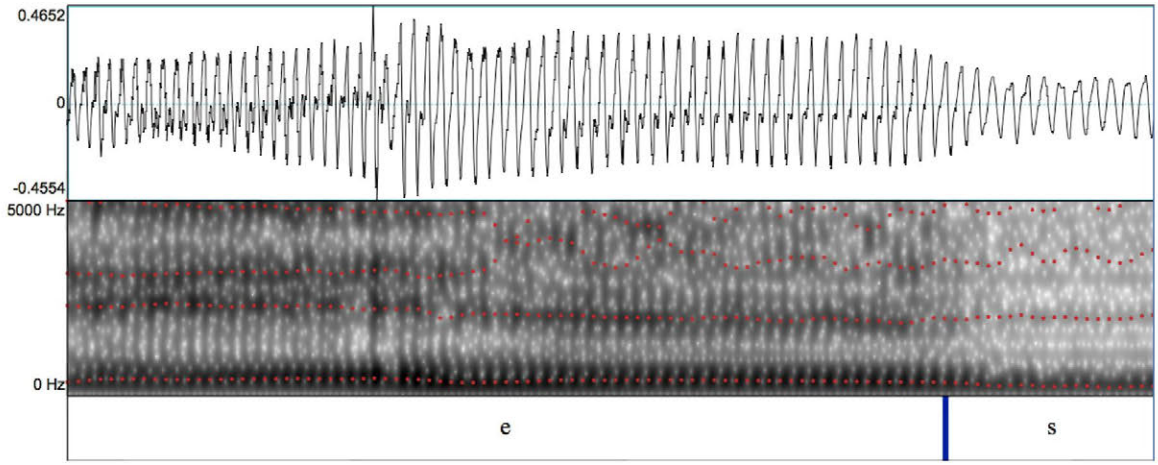
Fig. 1. Waveforms and spectrograms of a parkinsonian (newly diagnosed) and a control speaker pronouncing the syllable /pe/. Obtained from Neurovoz corpus, described in Section 3. Red dot lines mark the first four formants calculated with the software Praat [23].

and utterance aligning.[3] In the first iteration, the utterance is subdivided into a number of chunks equal to the number of monophones of the utterance, obtained from the transcription. Each chunk has the same time length in the first iteration but after a re-aligning using HMM and Viterbi algorithms, new segments are obtained and used to train a new GMM. This process is repeated for a certain number of iterations or until there is not substantial improvement in the alignment and a convergence is reached.

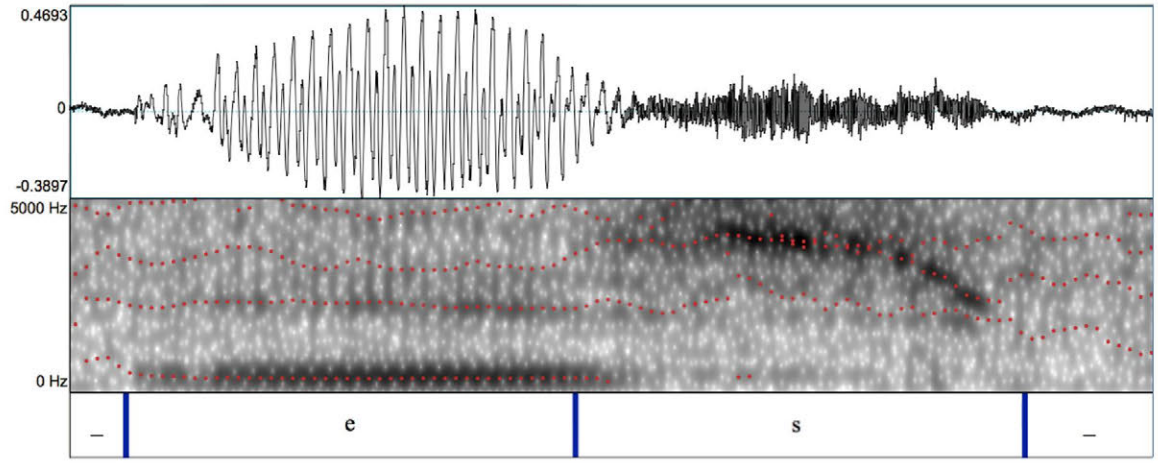
3. Using this monophone basic alignment as a basis, a new model is created following the same procedure but using triphones as phonetic units. Triphones are phonetic units that depend on the preceding and following phonemes. Thus, triphones help modelling the coarticulation effect. This new triphone GMM is optimized after several iterations, as in the previous step.
4. Once this GMM set is obtained and all the signals are aligned employing triphones, all the features are transformed using the Linear Discriminant analysis + Maximum Likelihood Linear Transform (LDA+MLLT) projection. The new features are used in a new sequence of training-alignment iterations considering triphones.
5. Finally, MFCC features are transformed using Feature space Maximum Likelihood Linear Regression (fMLLR) to obtain speaker independent features; and the last iterations of the triphone training-alignment process are performed.

Once a FAM is obtained, it can be applied to an utterance - characterized by the same features employed in the training (i. e., MFCC+Δ + $\Delta\Delta$)- and its associated transcription to automatically segment and label it. It is important to remark that the MFCC characterization is used only to train the FAM and to obtain the final labels but in this work, this feature family is not used directly to detect PD. The Kaldi FAM model, described previously, provides the phonetic labels (e. g, /p/, /a/, or /n/) for a certain utterance, without making any distinction between the different types of possible allophones that can be associated to a single phoneme. Only the plosive-fricative allophones of the phonemes /b/, /d/ and /g/ were considered and the use of the vowel /u/ separately and in diphthongs and hiatus, represented with the label /w/. Table 10 in Appendix A includes the correspondence between the different allophones of Castilian Spanish and the Kaldi phonetic labels.

[3] Alignment is referred to segmentation and labeling of the speech at the same time. This segmentation and labeling is used to iteratively train the new GMM more precisely after each iteration.



(a) Idiopathic PD female speaker. Age: 85. UPDRS: 47. Span: 331 ms



(b) Control female speaker. Age: 83. Span: 338 ms

Fig. 2. Waveforms and spectrograms of a parkinsonian (in an advanced stage) and a control speaker pronouncing the word "es". Obtained from Neurovoz corpus. Red dot lines mark the first four formants calculated with the software Praat.

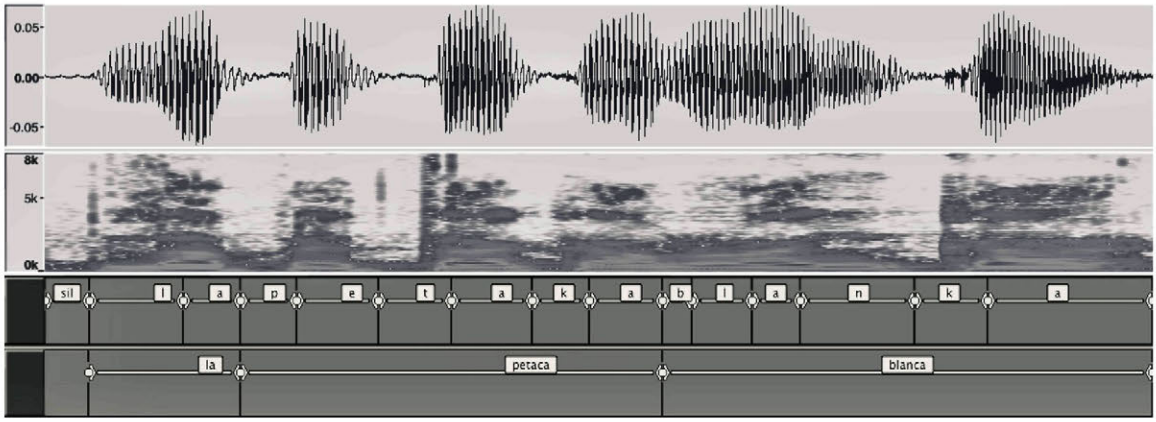


Fig. 3. Example of forced alignment labeling obtained with Kaldi, including waveform, spectrogram, phonemes and transcription.

2.2. Approach 1: forced Gaussian mixture model

One of the first techniques proposed in this work to detect PD from speech is the fGMM. This scheme trains GMMs where the Gaussian components of the mixture correspond to phonetic units instead of being obtained in an unsupervised procedure by maximum likelihood Expectation-Maximization (E-M) iterations [32]. To train the model, in the E-step, feature frames are aligned to phonetic units using the Kaldi FAM model and the Gaussian responsibilities are computed. In the M-step the parameters of GMM are updated by maximizing the EM objective in the usual way. The goal is to obtain improved GMMs (fGMM) which allow to compare more

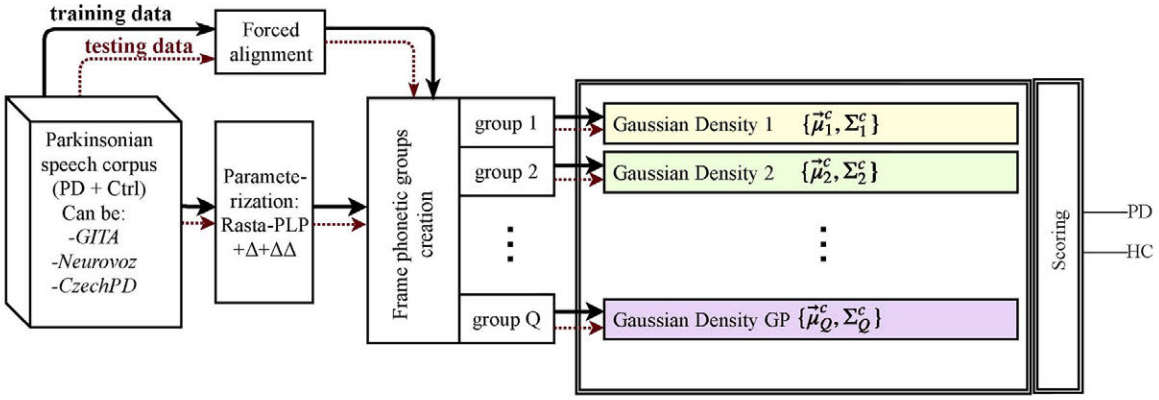


Fig. 4. fGMM general scheme. Initially, the fGMM model is trained using the training data (training fold) and then, tested with the testing data (testing fold). HC stands for healthy control.

precisely the differences between the realizations of each phonetic unit in each one of the two classes.

For a certain training corpus containing speakers from two classes, $\boldsymbol{\Upsilon}_c$ is the set of recordings of class c, composed of U_c utterances $\mathbf{X}_u$:

$$\boldsymbol{\Upsilon}_c = \{\mathbf{X}_1, \ldots, \mathbf{X}_u, \ldots, \mathbf{X}_{U_c}\} \tag{1}$$

Then, for an utterance containing N frames, the sequence of feature vectors, $\mathbf{X}_u$ is:

$$\mathbf{X}_u = \{\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N\} \tag{2}$$

where the vectors $\mathbf{x}_n$ have dimension D. For any utterance, let $\boldsymbol{\Xi}_u$ be the sequence of labels corresponding to $\mathbf{X}_u$:

$$\boldsymbol{\Xi}_u = \{\xi_1, \ldots, \xi_n, \ldots, \xi_N\} \tag{3}$$

where each ξ_n is the phonetic label of feature vector $\mathbf{x}_n$, obtained by the FAM. The number of different types of labels is Q. In a general case,

$$\xi_n \in \{\iota_1, \ldots, \iota_g, \ldots, \iota_Q\} \tag{4}$$

being ι_g the g^{th} phonetic label type (/h/, for instance). Then, the fGMM is represented by $\boldsymbol{\Theta}^c = \{\boldsymbol{\mu}_g^c, \boldsymbol{\Sigma}_g^c\}|_{g=1}^Q$, where $\boldsymbol{\mu}_g^c$ is the mean vector and $\boldsymbol{\Sigma}_g^c$ is the covariance matrix of Gaussian g in class c. In this work only the diagonal covariance matrices were used. To calculate the g^{th} mean vector and covariance matrix, only the frames from the g^{th} phonetic unit were used. For instance, to compute $\boldsymbol{\mu}_1^c$ and $\boldsymbol{\Sigma}_1^c$, only the feature vectors of class c labeled with ι_1, (normally /a/), were used. The result is one Gaussian per phonetic unit, with distribution $p_g^c(\mathbf{x}_n)$:

$$p(\mathbf{x}_n \mid H_c) = p_g^c(\mathbf{x}_n) \quad , \tag{5}$$

being H_c the hypothesis of a certain utterance belonging to class c.

In this case, there is a single g for each $\mathbf{x}_n$, determined by the phonetic label associated to this feature vector. For instance, if the feature vector $\mathbf{x}_n$ is labeled with /a/ ($g = 1$), then $p(\mathbf{x}_n \mid H_c) = p_1^c(\mathbf{x}_n)$. Note that in the fGMM, there is no soft Gaussian responsibilities as those used in the typical GMM models [32]. Since every single testing feature vector is unambiguously associated to a specific Gaussian component $p_g^c(\mathbf{x})$ (characterized by its mean vector $\boldsymbol{\mu}_g^c$ and covariance matrix $\boldsymbol{\Sigma}_g^c$), there is no need to use weightings to balance the use of the different $p_g^c(\mathbf{x})$.

For any feature vector $\mathbf{x}_n$ the Gaussian density $p_g^c(\mathbf{x}_n)$ is defined as:

$$p_g^c(\mathbf{x}_n) = \frac{\exp\{-1/2(\mathbf{x}_n - \boldsymbol{\mu}_g^c)'(\Sigma_g^c)^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_g^c)\}}{(2\pi)^{D/2}|\Sigma_g^c|^{1/2}} \tag{6}$$

Finally, the scores for each utterance $\mathbf{X}_u$ for the model of class c, containing N frames, are calculated by means of the log-likelihood of every frame as:

$$\Lambda_u^c = \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\Theta}^c) \quad , \tag{7}$$

and the global scores for each utterance are:

$$\Lambda_u = \Lambda_u^{\text{PD}} - \Lambda_u^{\text{control}} \quad . \tag{8}$$

Once the different scores are calculated for all vectors, to compute the class membership of a certain feature vector $\mathbf{x}_n$ from the test set, its score is compared with a threshold, λ to prove the hypothesis H_{PD}:

$$\Lambda_u \begin{cases} \geq \lambda, & \text{accept } H_{\text{PD}} \\ < \lambda, & \text{reject } H_{\text{PD}} \end{cases} \tag{9}$$

Fig. 4 contains a diagram of the training and testing processes of the fGMM scheme.

2.3. Approach 2: gaussian mixture model – forced universal background model

The GMM – forced Universal Background Model (GMM-fUBM) classification scheme is similar to the GMM – Universal Background Model (UBM) explained in [32]. However in this case, the UBM is created using the phonetic labels from the UBM speech corpus, obtaining a fUBM as explained in Section 2.2. The goal is to produce a UBM containing Gaussians directly associated to the various phonetic labels. Afterwards, the GMM-fUBM model for each class is obtained by performing Maximum a Posteriori (MAP) adaptation of the fUBM mean vectors to the adaptation-training utterances. In the adaptation step, the phonetic labels of the adaptation-training corpus (parkinsonian corpora, usually) are not used and frame-Gaussian alignment is performed using the Gaussian posterior probabilities.

Thus, after obtaining the fUBM model $\boldsymbol{\Theta}_{\text{fUBM}} = \{\boldsymbol{\pi}_g^{\text{UBM}}, \boldsymbol{\mu}_g^{\text{UBM}}, \boldsymbol{\Sigma}_g^{\text{UBM}}\}|_{g=1}^Q$, the GMM-fUBM model will contain the fUBM weights and covariance matrix and the MAP adaptation of the fUBM mean vectors, $\hat{\boldsymbol{\mu}}_g$, resulting in $\boldsymbol{\Theta}_f^c = \{\boldsymbol{\pi}_g^{\text{UBM}}, \hat{\boldsymbol{\mu}}_g^c, \boldsymbol{\Sigma}_g^{\text{UBM}}\}|_{g=1}^Q$.

Fig. 5 contains a diagram of the training and testing processes of the GMM-fUBM scheme.

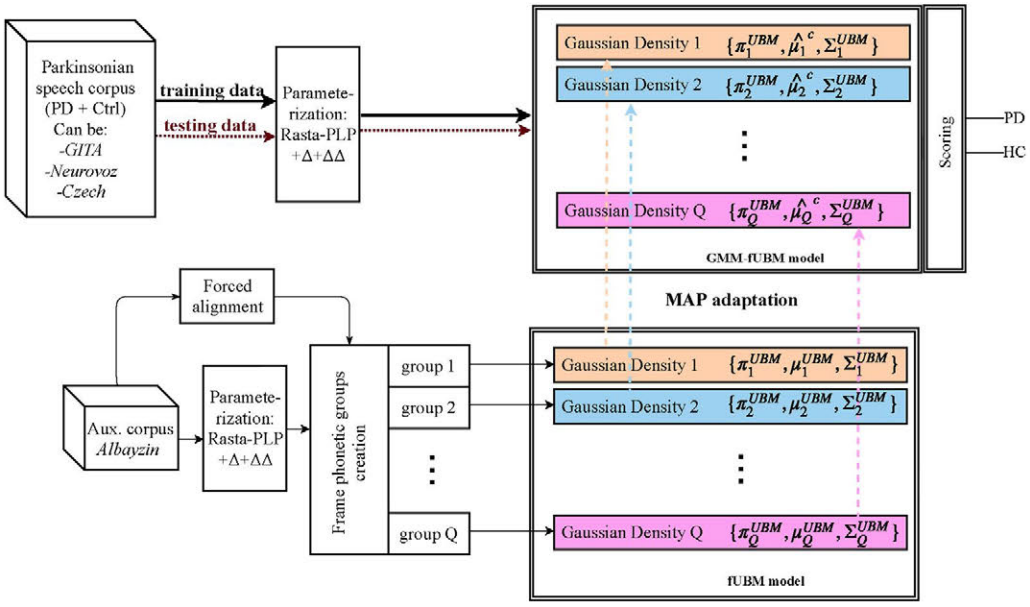


Fig. 5. GMM-fUBM general scheme. Initially, the model is trained using the training data (training fold) and then, tested with the testing data (testing fold).

2.4. Approach 3: forced gaussian mixture model – forced Universal Background Model

The the forced GMM – forced UBM (fGMM-fUBM) of each class are obtained by adapting mean vectors of the fUBM, but in this case force alignment with the phonetic labels of the adaptation-training corpus must be used. The goal of this scheme is that each of the resulting mixture densities is obtained using exclusively segments of a certain phonetic type (/a/, /b/, /d/, etc.). In this manner we can compare more precisely these types between classes, as in the fGMM but adapting the final models from a fUBM which allows to generalize better.

Similarly to the previous section, after obtaining the fUBM model $\boldsymbol{\Theta}_{\text{fUBM}} = \{\boldsymbol{\mu}_g^{\text{UBM}}, \boldsymbol{\Sigma}_g^{\text{UBM}}\}|_{g=1}^{Q}$, the fGMM-fUBM model will be composed of the fUBM covariance matrix and the adaptation of the fUBM mean vectors using the adaptation-training corpus and its correspondent phonetic labeling, resulting in $\boldsymbol{\Theta}_{f-f}^{c} = \{\hat{\boldsymbol{\mu}}_{\text{f-f},g}^{c}, \boldsymbol{\Sigma}_g^{\text{UBM}}\}|_{g=1}^{Q}$.

The adapted means are calculated as

$$\hat{\boldsymbol{\mu}}_{\text{f-f},g}^{c} = \alpha_g^c \boldsymbol{\mu}_g^c + (1 - \alpha_g^c)\boldsymbol{\mu}_g^{\text{UBM}} \tag{10}$$

where

$$\alpha_g^c = \frac{r_g^c}{r_g^c + \sigma} \tag{11}$$

being r_g^c the number of frames of class c in the training corpus labeled as ι_g and σ, is the relevance factor, as defined in [32].

Again, to compute the class membership of a certain feature vector $\mathbf{x}_n$ from the testing subgroup, the procedure explained in Section 2.2 is applied.

Fig. 6 contains a diagram of the training and testing processes of the fGMM-fUBM scheme.

2.5. Approach 4: optimized forced gaussian mixture model – forced universal background model

Taking advantage of the phonetic labeling of the utterances and the unambiguous correspondence between the speech segments and the Gaussian mixtures, a modification of the previous scheme is proposed to optimize the models during the training: the optimized- forced GMM - forced UBM (opt-fGMM-fUBM). The idea behind this optimization is to recursively compute the accuracy of each of the phonetic types of frames after obtaining the fGMM-fUBM classifier, using this accuracy to change the weight of each frame type in the computation of the class membership. Given the limited amount of available data, instead of using a held out set, the accuracy for each phonetic type was evaluated on the training set. Thus, if a phonetic group provides better accuracy than others on the training subset, the weighting vector will make the frames of that phonetic type more relevant in the final computation of the score for a new test utterance.

Therefore, after creating the fGMM-fUBM models, the membership of the training feature vectors to the two possible classes is computed and this information is used to create the weighting vector Φ^c,

$$\Phi^c \in \{\phi_1^c, \ldots, \phi_g^c, \ldots, \phi_Q^c\} \tag{12}$$

where

$$\phi_g^c = \frac{\text{nf}_{rd,g}^c}{\text{nf}_{t,g}^c} \quad , \tag{13}$$

being $nf_{rd,g}^c$ the number of frames of phonetic group g and class c correctly classified, and $nf_{t,g}^c$ the total number of frames of phonetic group g and class c.

In this case, the new scores of any feature vector $\mathbf{x}_n$ respect to a certain model of class c will be

$$\Lambda_{\text{opt-f-f},n}^{c} = \log\left(p(\mathbf{x}_n|\boldsymbol{\Theta}_{\text{f-f}}^{c})\right) \sum_{g=1}^{Q} \phi_g^c \eta_{g,n} \quad , \tag{14}$$

being $\eta_{g,n}$ an index indicating if the phonetic label associated to $\mathbf{x}_n$ is ι_g ($\eta_{g,n} = 1$) or not ($\eta_{g,n} = 0$). Therefore, the global scores for a certain utterance of N frames are

$$\Lambda_{\text{opt-f-f}} = \frac{1}{N}\sum_{n=1}^{N}\left(\Lambda_{\text{opt-f-f},n}^{\text{PD}} - \Lambda_{\text{opt-f-f},n}^{\text{control}}\right) \tag{15}$$

The weights Φ^c are used to evaluate the performance of the models and the obtained results yield new weights. The process is repeated until reaching a maximum number of iterations or until the overall training accuracy does not increase more than a certain incremental improvement, ϖ_{min}.

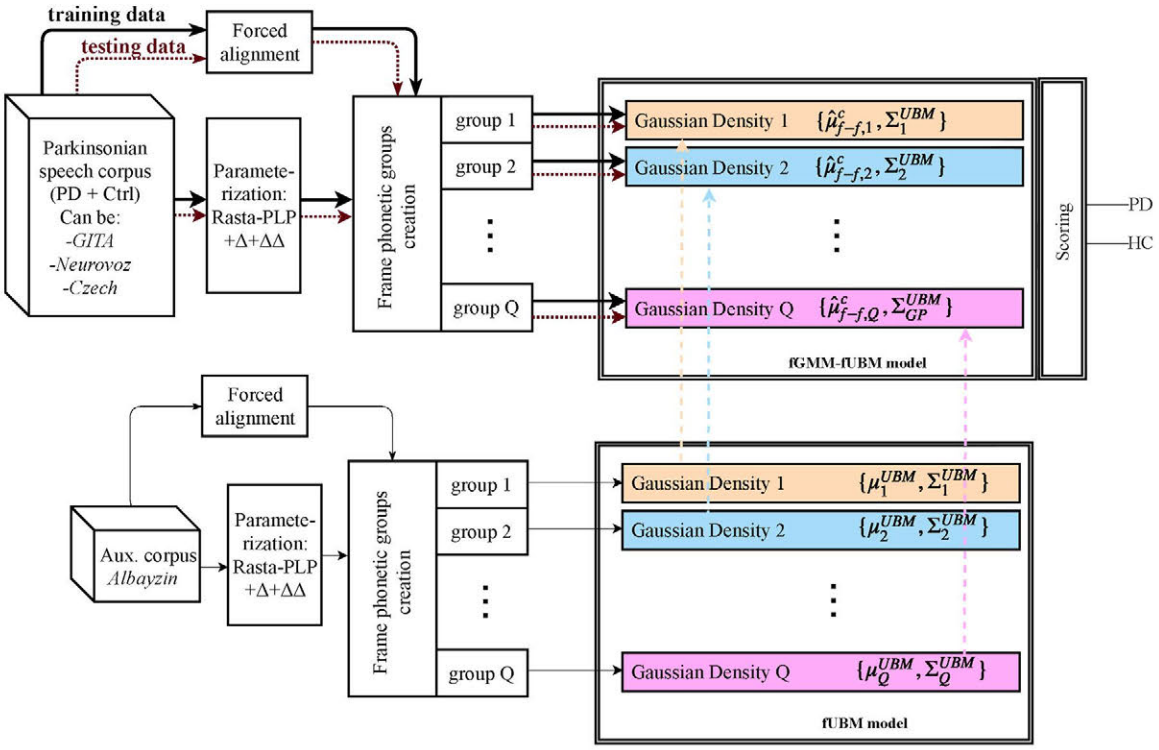


Fig. 6. fGMM-fUBM general scheme. Initially, the model is trained using the training data (training fold) and then, tested with the testing data (testing fold).

3. Experimental setup

3.1. Materials: speech corpora

In this work, five corpora were employed: Neurovoz, GITA, CzechPD, Albayzin and FisherSP. The first three contain different speech tasks from PD patients and control speakers. These were used to train, adapt and test the different classification models depending on the proposed approaches. On the other hand, Albayzin and FisherSP are two auxiliary corpora in Spanish language used to train the UBM and FAM respectively.

Table 1
Demographic statistics of Neurovoz corpus. Ages are expressed in years.

	Female		Male	
	PD	HC	PD	HC
#Subjects	18	18	29	14
Age, average	70.9	68.4	71.9	66.6
Age range	59–86	58–83	41–88	55–77
UPDRS, average	18.2	–	7.4	–
H&Y, average	2.3	–	2.3	–
Years since diagnosis	6.6	–	7.4	–

3.1.1. Neurovoz

Neurovoz is a new corpus containing 47 parkinsonian and 32 control speakers whose mother tongue is Castilian Spanish. This corpus was recorded in collaboration with the otorhinoaringology and neurology services of the Gregorio Marañón hospital in Madrid, Spain. The recruitment of patients and the recording of the corpus was approved by the Ethics Committee of the hospital. The sub-set employed in this work contains a Diadochokinetik (DDK) task (repetitions of the syllable sequence /pa-ta-ka/), six fixed sentences and running speech from the description of a picture. All of these tasks were produced at a comfortable speech Sound Pressure Level (SPL). Regarding the fixed sentences or text-dependent utterances, these were first listened by the subjects from a recording of a standard speaker and then repeated, instead of read from a text document. This procedure diminishes the noise of paper during recording, the reading mistakes caused by vision problems common in elderly and the cognitive load of the process.

Table 1 shows the age, sex and severity statistics of subjects in the corpus while Table 2 includes the six fixed sentences, its phonetic transcription and its translation to English. The UPDRS, Hoehn & Yahr rating scale (H&Y) and years since diagnosis distributions of the PD patients are portrayed in Fig. 7.

All of the patients were under pharmacological treatment and took the medication between 2 and 5 h before the speech and voice recording. Their neurological state was assessed by a neurologist right before the recording session. After this first examination, an otolaryngologist and a speech therapist evaluated perceptually and objectively the patients' voice. An assessment of the phonatory system through anamnesis and visual examination of the vocal folds with a fiber-laryngoscope was carried out to discard organic pathologies.[4] A survey was conducted with the control group to assess their neurological state. A speech therapist assessed their voice perceptually and by means of a survey. Speakers with organic or neurological pathologies (other than PD) were discarded, as well as smokers and subjects with alcoholic addictions. Recordings of both groups were performed in the same room with controlled acoustic characteristics.

The transducer used to record this corpus is an AKG C420 headset microphone which was coupled to a preamplifier with phantom power. The signal from the preamplifier was routed to a Soundblaster Live 24 bits sound card connected to a personal computer equipped with the software Medivoz [33]. The sampling rate was 44.100 Hz, and the quantization, 16 bits.

[4] 11 patients refused to do the fiber-laryngoscope exploration and were assessed only perceptually and through anamnesis.

Table 2
Spanish transcription of the six Neurovoz fixed sentences and its translation to English.

Sentence #	Spanish transcription/Phonetic transcription/*English translation*
1	Cuando las barbas de tu vecino veas pelar, pon las tuyas a remojar / 'kwaŋdo laz 'β ar 'β az ð e tu 'β e'θ ino 'β eas pe'lar 'pon las 'tuʝ as a remo'xar / *When the beard of your neighbor you see peel, put yours to soak*
2	De la calle vendrá quien de tu casa te echará / de la 'kaʎ e 'β eŋ'd r a 'kjeŋ de tu 'kasa te etʃ a' r a / *From outside will come who from your house will kick you out*
3	Cuando el diablo no sabe qué hacer, con el rabo mata moscas / 'kwaŋdo el 'dja'β lo no 'sa'β e 'ke a'θ e r 'kon el 'ra'β o 'mata 'moskas / *When the devil does not know what to do, kills flies with the tail*
4	La petaca blanca es mía / la pe'taca 'β laŋka ez 'mia / *The white flask is mine*
5	No pidas a quien pidió, ni sirvas a quien sirvió / no 'pið as a 'kjem pi'ð jo ni 'si r bas a 'kjen si r 'bjo / *Do not beg the one who begged, nor serve the person who served*
6	El que a buen árbol se arrima, buena sombra le cobija / el ke a 'β wen 'a r 'β ol se a'rima 'bwena 'somb r a le ko'β ixa / *To the one that comes to a good tree, good shade covers him*

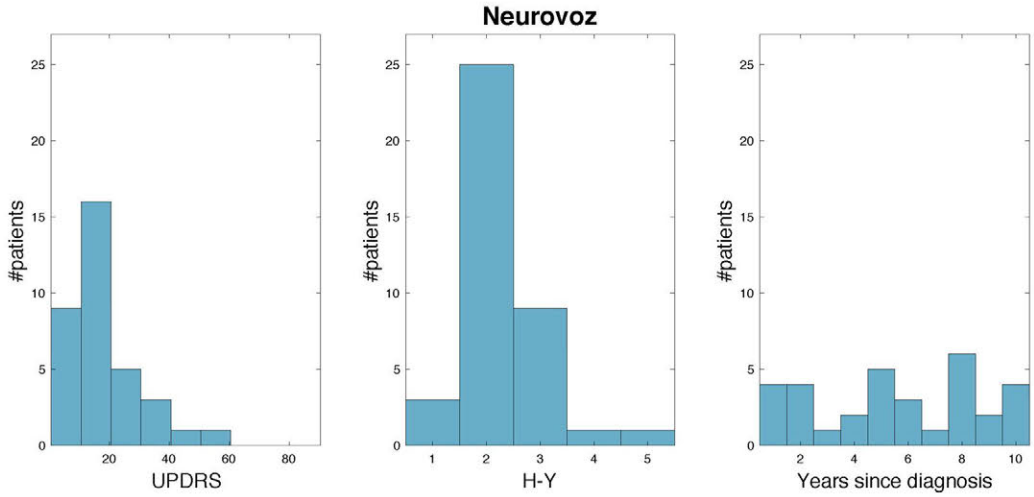


Fig. 7. UPDRS, H&Y and years since diagnosis histograms for the PD patients within the Neurovoz corpus.

Table 3
Demographic statistics of GITA database. Ages are expressed in years.

	Female		Male	
	PD	HC	PD	HC
#Subjects	25	25	25	25
Age, average	60.7	61.4	61.6	60.5
Age range	49–75	49–76	33–81	31–86
UPDRS, average	37.6	–	37.7	–
H-Y, average	2.2	–	2.3	–
Years since diagnosis	12.6	–	8.9	–

Table 4
Demographic statistics of CzechPD database. Ages are expressed in years.

	Male	
	PD	HC
#Subjects	20	16
Age, average	61.0	61.8
Age range	34–83	36–80
UPDRS, average	17.9	–
H-Y, average	2.1	–
Years since diagnosis	2.38333	–

3.1.2. GITA

GITA is a Colombian corpus presented in [34], containing a variety of speech tasks from 50 patients with PD and 50 age- and sex-matched control speakers whose native language is Colombian Spanish.

Three types of speech tasks from the GITA corpus were used in this work, comprising a DDK task (/pa-ta-ka/), a monologue and six text-dependent tasks (read sentences).

Table 3 shows the demographic statistics of GITA while Fig. 8 portrays the UPDRS, H-Y and years since diagnosis distributions of the PD patients.

A more detailed description of GITA can be found in [34].

3.1.3. CzechPD

The CzechPD corpus employed in this work contains a DDK task /pa-ta-ka/ from 20 newly diagnosed and untreated parkinsonian speakers and 14 control speakers whose mother tongue is Czech. This dataset is described in detail in [18]. Unlike the other two parkinsonian corpora, none of the patients from CzechPD was under treatment.

Table 4 shows the age, sex and severity statistics of subjects in the corpus while the UPDRS, H&Y and years since diagnosis distributions of the PD patients are portrayed in Fig. 9.

3.1.4. Auxiliary corpora

The Albayzin corpus [35] is a phonetically balanced dataset, sampled at 16 kHz and quantized with 16 bits, composed by a large amount of utterances in Spanish language and their transcriptions. For this work purposes, only the first subset from the five provided in the corpus is used to obtain the UBM.

On the other hand, the FisherSP (Fisher Spanish) corpus has been created by the Linguistic Data Consortium to develop auto-

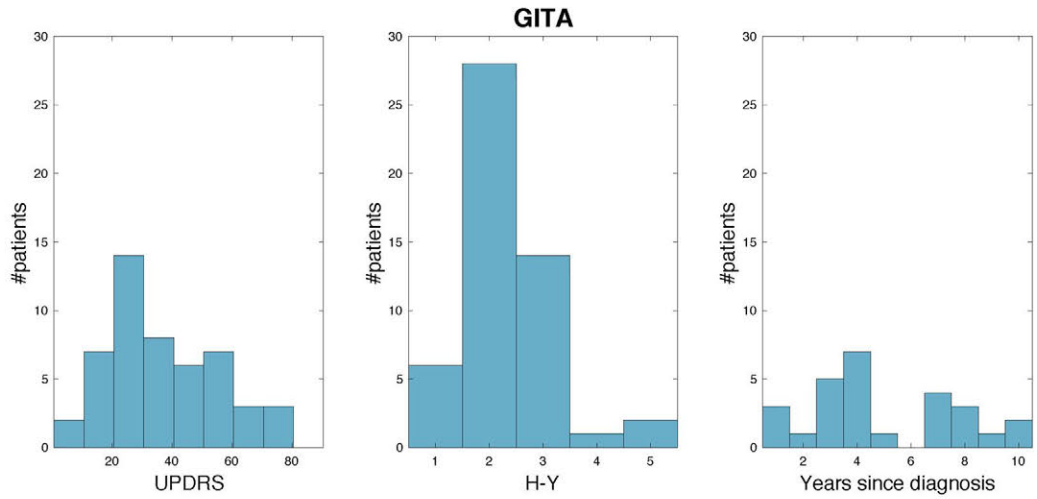


Fig. 8. UPDRS, H-Y and years since diagnosis histograms for the PD patients within the GITA database.

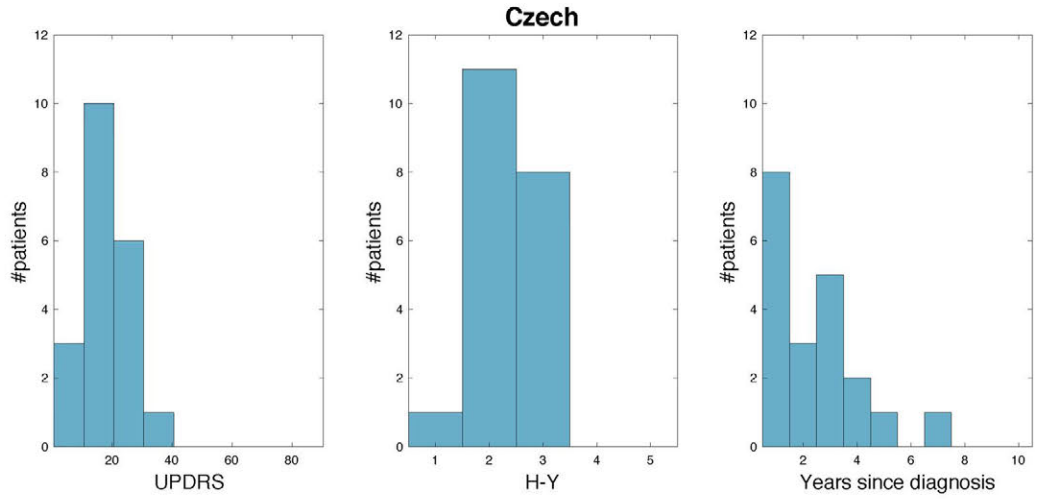


Fig. 9. UPDRS, H-Y and Years since diagnosis histograms for the PD patients within the CzechPD database.

matic speech recognizers in Spanish language. It contains around 163 hours of telephonic speech from 136 native Spanish language speakers from at least 20 countries but mostly from the USA, along with their transcriptions.[5] This database is sampled at 8 kHz and 16 bits. In this work, it is used to train the FAM.

3.2. Methodology

The purpose of the methodology of this experimental set is to analyze the four classification approaches exposed in Section 2, for the automatic detection of PD from speech. To carry out these tests, it is necessary to perform beforehand the forced alignment of GITA, Neurovoz and Albayzin corpora.

3.2.1. Forced alignment

The FisherSP corpus was used to train a FAM employing Kaldi. This corpus, sampled at 8 kHz is normalized and divided into frames of 25 ms with a frame shift of 10 ms. A vector containing 20 MFCC+Δ+$\Delta\Delta$ features was processed from each frame.[6] To carry out the initial monophone training-alignment, a subset of 10,000 utterances was randomly selected, producing the monophone model (40 iterations). After this, an alignment was performed in a subset of 30, 000 utterances to obtain new segments for the second round of training-alignment. Afterwards, two rounds of 30 iterations were performed to train a triphone model using 30, 000 utterances followed by the correspondent alignments. Then, a round of 35 iterations employing a LDA+MLLT transformation of the acoustic features as input was performed using 100, 000 utterances. Finally, 30 iterations of the training-aligning sequence using fMLLR transformation of the features produced the final FAM.

This FAM was used to segment and label all the 6 text-dependent tasks of the GITA corpus, 6 text-dependent tasks from Neurovoz corpus and all the recordings of the training subset of the corpus 1 from Albayzin (4.801 fixed sentences), employing in all the cases their correspondent transcriptions within the process. For these purposes, all of these corpora were normalized, filtered and downsampled to 8 kHz and characterized with 20 MFCC+Δ+$\Delta\Delta$ features per frame (25 ms, 10 ms frame shift), as the corpus utilized to train the FAM.

3.2.2. Forced gaussian mixture models

The four approaches described in Section 2 were assessed for the detection of PD employing mainly GITA and Neurovoz separately to

[5] Full description and transcriptions available at https://catalog.ldc.upenn.edu/ldc2010t04

[6] These parameter values and set-up are employed following the recommendations of the used Kaldi recipe available at the platform repository.

Table 5
fGMM results.

Corpus	Speech task	Accuracy ± CI (%)	AUC	Sens.	Spec.	N	UPDRS ρ	p-Val.	H&Y ρ	p-Val.
GITA	Text-dependent utterances	79 ± 8	0.88	0.76	0.82	10	0.46	0.001	0.27	0.067
Neurovoz	Text-dependent utterances	71 ± 10	0.86	0.67	0.77	14	0.47	0.004	0.32	0.049

train or adapt the models and Albayzin to generate the UBM. As no FAM was generated for Czech language, only the second approach (GMM-fUBM) was evaluated with CzechPD. The validation of the models was carried out using a k-fold cross-validation scheme, with $k = 11$. Finally, several cross-corpora trials were performed using the three parkinsonian corpora.

Thus, a first round of trials was carried out to train and test fGMM models (approach 1) with GITA and Neurovoz corpora separately. In the second round, GMM-fUBM models (approach 2) were trained and tested separately for GITA, Neurovoz and CzechPD, using Albayzin for the UBM. In the third and fourth rounds, fGMM-fUBM (approach 3) and opt-fGMM-fUBM (approach 4) models respectively, were trained and tested separately using in this case GITA and Neurovoz corpora, employing Albayzin to create the UBM. For the opt-fGMM-fUBM, the maximum number of iterations was limited to 20 and $\varpi_{min} = 0.5\%$. In all the cases, the recordings were parameterized using Rasta-Perceptual Linear Predictive (Rasta-PLP) $+\Delta+\Delta\Delta$ with a number of PLP coefficients, F, varying in the range $\{10 \ldots 20\}$ with steps of 2. All the signals from the parkinsonian databases were filtered and downsampled to 16 kHz when necessary, to match the sampling frequency of Albayzin. The used frame length was 15 ms and the number of coefficients in the Finite Impulse Response (FIR) filter, 5. This set-up was selected on the basis of the results obtained in a previous work employing speaker recognition technologies to detect PD [36].

Respecting the speech tasks, all the available text-dependent utterances were employed jointly to train the respective GITA and Neurovoz models in all the trials. Additionally, in the second round of trials, the DDK task and monologues were utilized too. Therefore, for each round of trials 11 models were trained and tested per corpus, speech task and F. It is important to remark that the Rasta-PLP $+\Delta+\Delta\Delta$ coefficients were calculated for the whole utterance and, after obtaining the feature vectors, these are distributed into the correspondent phonetic bins for training or testing, attending to their forced alignment labeling. In the k-folds trials, none of the speakers used for training or adaptation were used for testing.

Finally, a round of cross-corpora trials is performed using the GMM-fUBM scheme only with DDK tasks from GITA, Neurovoz and CzechPD. These trials consist in the combination of two of the parkinsonian corpora to train a model which is tested with the remaining corpus. This process is repeated for all the possible combinations. DDK tasks are employed since these are the only common speech tasks in the three corpora.

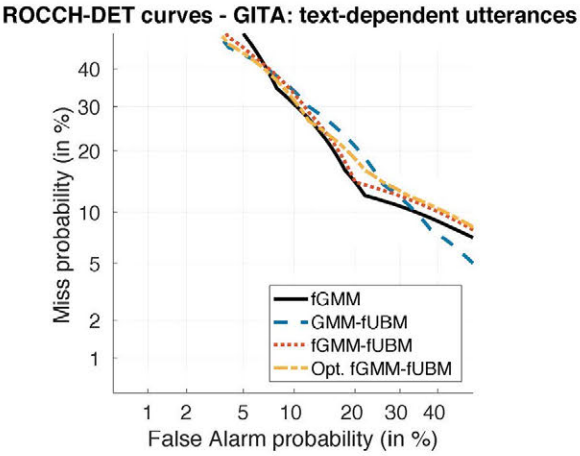


Fig. 10. Convex hull around receiver operating characteristic – detection error trade-off (ROCCH-DET) curves obtained employing testing scores from GITA in the classification schemes: fGMM, GMM-fUBM, fGMM-fUBM and opt-fGMM-fUBM. Text-dependent utterances are used as speech task.

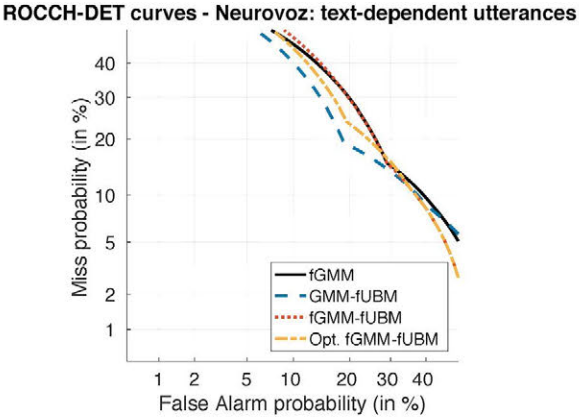


Fig. 11. ROCCH-DET curves obtained employing testing scores from Neurovoz in the classification schemes: fGMM, GMM-fUBM, fGMM-fUBM and opt-fGMM-fUBM. Text-dependent utterances are used as speech task.

4. Results

In this section, the results from all the rounds of trials are detailed. All of the tables and figures are referred to the k-folds cross-validation unless specified.

Table 5 shows the results from the analysis of the fGMM scheme, Table 6 shows the results related to the GMM-fUBM models, while Tables 7 and 8 include the results referred to the fGMM-fUBM and opt-fGMM-fUBM classification schemes, respectively. All of these tables include the classification accuracy ± Confidence Interval (CI), Area Under the ROC Curve (AUC), sensitivity, specificity, N, and the Spearman's correlation (ρ) of the scores with the UPDRS and H&Y scales with their respective p-values. In all the cases, the CI was calculated as indicated in [37]. Best results for each corpus are in bold.

Figs. 10 and 11 plot the ROCCH-DET curves for the GITA and Neurovoz corpora respectively, using the four proposed classification schemes.

Figs. 12 and 13 show the accuracy per frame phonetic group in GITA and Neurovoz corpora respectively, employing text-dependent utterances in an opt-fGMM-fUBM scheme. In these figures, the phonetic labels B and D are referred to fricative allophones while b and d are plosives. R is multiple vibrant and r is simple vibrant.

Table 9 includes the results of the cross-corpora tests using GITA, Neurovoz and CzechPD as testing corpora separately. For each case, the other two parkinsonian corpora are used to train the model. Lastly, Fig. 14 plots the ROCCH-DET curves in the cross-corpora trials using the DDK task in a GMM-fUBM classification scheme.

Table 6
GMM-fUBM results.

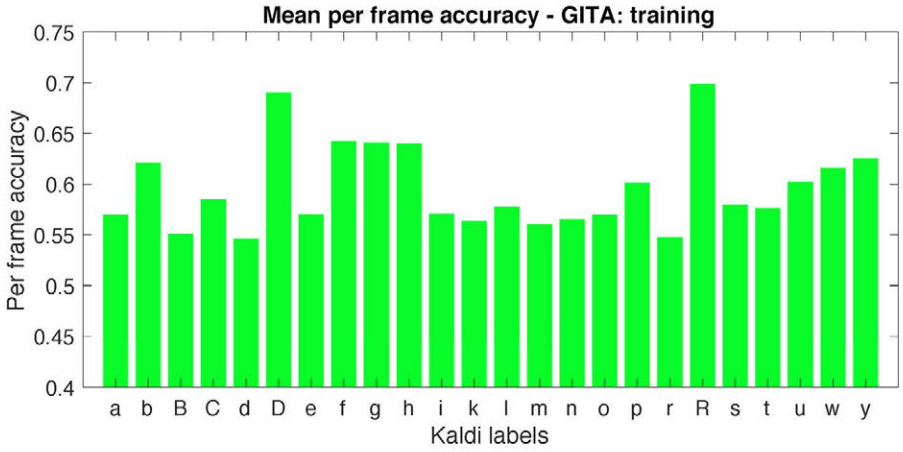
Corpus	Speech task	Accuracy ± CI (%)	AUC	Sens.	Spec.	N	UPDRS ρ	p-Val.	H&Y ρ	p-Val.
GITA	Text-dependent utterances	78 ± 8	0.88	0.78	0.78	12	0.18	0.238	0.10	0.423
	DDK	79 ± 8	0.86	0.86	0.72	12	0.20	0.188	0.19	0.196
	Monol.	78 ± 8	0.84	0.73	0.82	10	0.43	0.002	0.28	0.054
Neurovoz	Text-dependent utterances	**81 ± 9**	**0.87**	**0.83**	0.78	14	0.21	0.232	−0.05	0.762
	DDK	81 ± 9	0.85	0.83	0.77	20	0.31	0.070	0.13	0.422
	Monol.	66 ± 14	0.67	0.35	0.83	10	0.47	0.109	0.17	0.586
CzechPD	DDK	**94 ± 6**	**0.97**	**0.9**	1	20	0.06	0.808	−0.05	0.838

Table 7
fGMM-fUBM results.

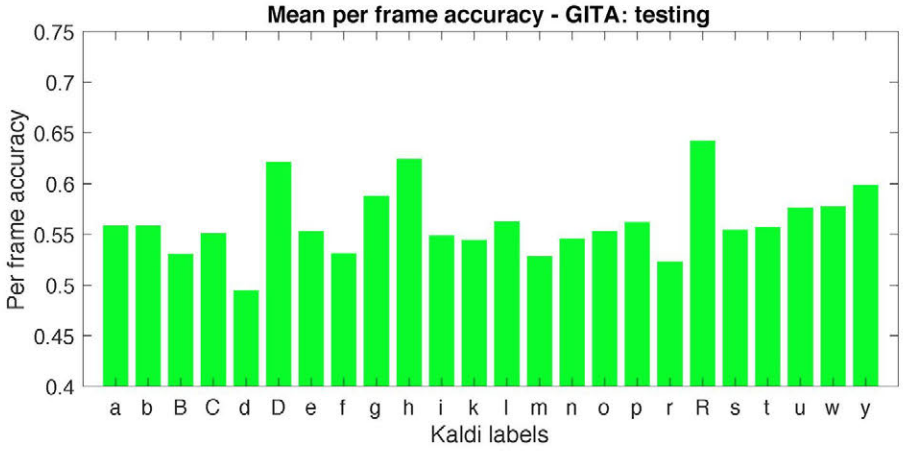
Corpus	Speech task	Accuracy ± CI (%)	AUC	Sens.	Spec.	N	UPDRS ρ	p-Val.	H&Y ρ	p-Val.
GITA	Text-dependent utterances	**81 ± 8**	**0.88**	**0.84**	0.78	10	0.40	0.006	0.31	0.038
Neurovoz	Text-dependent utterances	75 ± 10	0.86	0.76	0.74	14	0.36	0.033	0.31	0.056

Table 8
Opt-fGMM-fUBM results.

Corpus	Speech task	Accuracy ± CI (%)	AUC	Sens.	Spec.	N	UPDRS ρ	p-Val.	H&Y ρ	p-Val.
GITA	Text-dependent utterances	**81 ± 8**	**0.88**	**0.84**	0.78	10	0.40	0.005	0.30	0.041
Neurovoz	Text-dependent utterances	79 ± 9	0.87	0.8	0.77	12	0.36	0.034	0.34	0.032

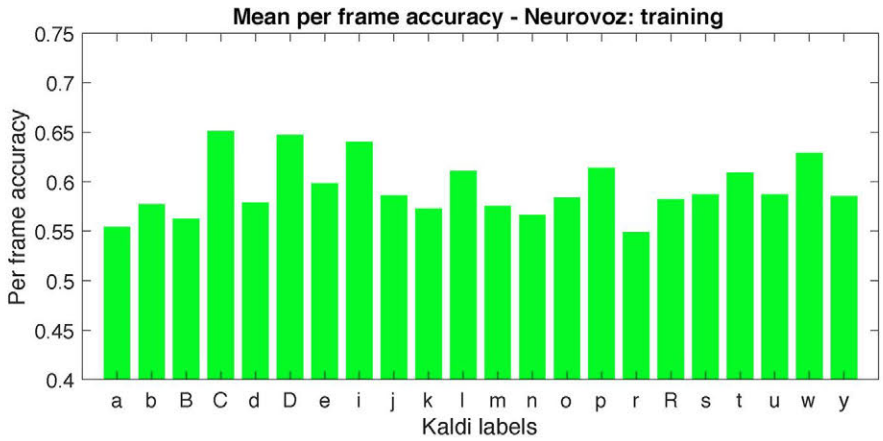


(a) Accuracy per frame phonetic group in GITA using training scores.

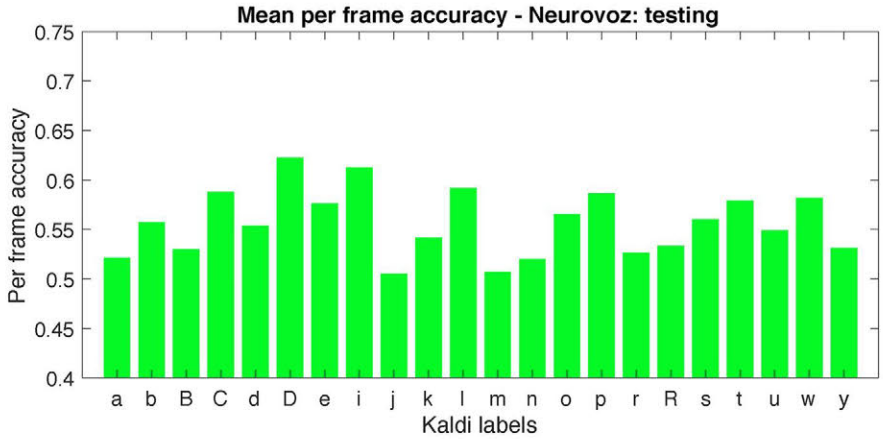


(b) Accuracy per frame phonetic group in GITA using testing scores.

Fig. 12. Per frame accuracy in GITA corpus using text-dependent utterances in an opt-fGMM-fUBM scheme. The horizontal line at the 0.5 of accuracy represents a random decision.



(a) Accuracy per frame phonetic group in Neurovoz using training scores.



(b) Accuracy per frame phonetic group in Neurovoz using testing scores.

Fig. 13. Per frame accuracy in Neurovoz corpus using text-dependent utterances in an opt-fGMM-fUBM scheme. The horizontal line at the 0.5 of accuracy represents a random decision.

Table 9
Cross-corpora GMM-fUBM results using DDK task. For each test corpus, the remaining two parkinsonian corpora are used to train the model.

Test corpus	Speech task	Accuracy ± CI (%)	AUC	Sens.	Spec.	N	UPDRS ρ	p-val.	H&Y ρ	p-val.
GITA	DDK	66 ± 9	0.76	0.9	0.42	18	0.22	0.1316	0.11	0.4767
Neurovoz	DDK	74 ± 10	0.78	0.87	0.5	10	0.2	0.2557	0.09	0.5822
CzechPD	DDK	76 ± 14	0.87	0.95	0.5	16	0.12	0.6066	0.07	0.7792

5. Discussion

In this experimental set, several approaches were proposed to create GMMs in such a manner that, after obtaining the final model, each Gaussian only represents a phonetic unit (for instance, the unit /a/). To obtain the Gaussian densities within the GMM, the different phonetic segments of the available speech corpora were employed, obtained with speech forced alignment techniques.

The approaches described in this experimental set are based on the use of GMM and GMM-UBM techniques but with differences in the material and in the iterative processes used to train or adapt the different models. Thus, the proposed classification schemes are: fGMM, GMM-fUBM, fGMM-fUBM and opt-fGMM-fUBM, described in Section 2.

5.1. *The glass ceiling in the automatic detection of PD from speech*

Although the minimum error found in all the analysed detectors in this work ranges from 6% to 19%, depending on the database under study, the theoretical limit of false rejection for this type of works is not clearly delimited and does not have to be 0% necessarily, like in other applications such as speaker recognition. To this respect, some specialists suggest that a 90% of PD patients suffer from dysarthria after a median latency period of 7 years since diagnosis [38,39]. These considerations would set the false rejection in automatic detectors employing speech to 10%. Recent works like [40] 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 [41] in a cohort of 48 patients in several stages. However, this work is focused in perceptual estimations or preliminary quality of voice analysis. No work has studied in detail the prevalence of the perturbations in voice and speech during the early stages of PD neither the differences between the dysphonia and dysarthria caused by PD and by aging. Thus, the minimum theoretical limit for false rejection in any study trying to detect PD from voice or speech could be higher than 0%. This value depends on the percentage of the patients included in the speech corpus with signs in voice or speech caused by PD and this

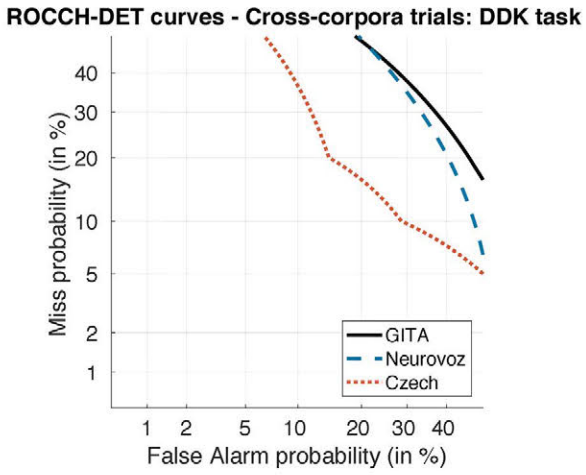


Fig. 14. ROCCH-DET curves obtained employing testing scores GITA, Neurovoz and CzechPD in the GMM-fUBM classification scheme. DDK sequence (/pa-ta-ka/) is used as speech task.

is dependent on the severity and stage of each patient. In the same manner, the percentage of patients without perceived dysarthria (most of them in early stages) but with true deviations in articulation is unknown. In the same manner, some studies point out that between 2% and 4% of patients diagnosed with idiopathic PD are really suffering from other type of pathology, as autopsies reveal [42,27]. Also, these values are obtained post-mortem and since the diagnosis is more precise as the disease progresses, these error rates will be higher for populations with patients in the early stages [43]. This means that some of the patients included in a parkinsonian speech corpus could not be really affected by PD, being this more likely in the newly diagnosed ones.

These two considerations would impose a minimum false rejection percentage to values which could be close to 10%.

Also, the use of non parkinsonian or non-dysarthric patients in the parkinsonian class within the training process can provide less accurate models to represent that specific class.

In the same manner, PD affects to 1% of population older than 60 and to 3% older than 80 according to recent studies [44]. That means that in works using 100 control speakers over 60 years, between 1 and 3 could be suffering from PD without being diagnosed, which will impose a minimum false acceptance rate around 2%.

All of these considerations limit the maximum achievable accuracy of the automatic detection systems in studies like the present work that, depending on the particularities of the corpora employed for a certain study, could be less than 95%.

5.2. Forced gaussians

All of the proposed schemes were analyzed employing separately GITA and Neurovoz, performing a *k*-folds cross-validation. Additionaly, the GMM-fUBM scheme was analyzed employing CzechPD too. Finally, the three parkinsonian corpora were employed in a cross-corpora validation round of trials.

For GITA, the opt-fGMM-fUBM approach provides the best results (accuracy equals to 81% and AUC equals to 0.88) while for Neurovoz the GMM-fUBM scheme is the one yielding the highest accuracies and AUC (accuracy equals to 81% and AUC equals to 0.87). For this last scheme, the accuracy with CzechPD is 94% and AUC, 0.97. In general, the CI intervals for the best results of GITA and Neurovoz are over ±9%, which produces an overlapping among the final result ranges. Therefore, the analysis of the results is focused on the observed trends.

It can be considered that the GMM-fUBM approach provides, in general lines, the best results since, for GITA and Neurovoz produces the best AUC and similar accuracies. In addition, this last scheme does not require to use forced alignment in the parkinsonian corpus, allowing the possibility to create supervised or unsupervised schemes, depending on the speech material used (text-dependent utterances and DDK will produce supervised approaches while monologues, unsupervised approaches).

Regarding the UPDRS correlation with the scores, the first approach, employing fGMM provides higher values for GITA and Neurovoz. It is important to observe that only the global UPDRS values were available but the study of the correlation with the motor part of UPDRS (part III) would have been more advisable. In all the cases, correlation with H&Y scale is lower. This can be caused by the fact that Spearman's correlation was employed instead of other type of methods that are more adequate to calculate correlation between continuous (scores) and discrete (ratings) values.

The cross-corpora results found in Table 9 demonstrate a reduced accuracy compared with the k-folds cross validations shown in Tables 5, 6–8.

5.3. Phonetic units

Observing Figs. 12 and 13 it is possible to deduce that the phonetic units leading to better accuracies using GITA are /D/, /g/, /h/ (aspirated h), /l/, /R/, /w/ and /y/ while /B/, /d/, /k/, /m/ and /r/ provide the lowest accuracies. Regarding Neurovoz, /C/, /D/, /i/, /l/, /p/ and /w/[7] produce the highest accuracies while /a/, /j/ and /m/, the lowest. In both corpora, phonetic units requiring a higher narrowing of the vocal tract but without a burst, such as the units /C/, /D/, /g/ and /R/ tend to be more decisive in the detection while others such as /B/ or /m/ are less influential. The impact of PD in fricatives has already been observed in other works such as [11–14]). Regarding vowels, /u/ and /w/ produce good results in both corpora, supporting the findings of [16] and [18].

The differences in the results of the GITA and Neurovoz trials are caused by the dialectal divergences between Castilian and Colombian Spanish and by the different position of the phonetic units within the sentences, which conditions its articulation (the text-dependent utterances are different in the two corpora). In general terms, Figs. 12 and 13 suggest that the differences in accuracy between the distinct specific phonetic units is variable and there is not a phonetic unit or groups of units that can clearly be used separately to detect PD. Also, this suggest that PD may affect to the articulatory sequence as a whole with a higher influence on phonetic units requiring a narrowing of the vocal tract but without a burst and with a lower influence in nasals. However, more general categories of articulatory movements related with the type of narrowing of the vocal tract or the use of the glottal source must be performed.

5.4. Other considerations

The proposed methodologies can be considered novel although previous works such as [45] used a similar approach for speaker recognition employing DNN posteriors instead of labels from speech forced alignment to create the UBM in an *i-Vectors* classification scheme. However, on spite of the similarities, it is the first time that this type of approach is used for the detection of PD.

The results obtained in this work are not higher than those obtained with GITA in [36] although promising results are obtained in the CzechPD trials where a 94% of accuracy is reached. One of the

[7] /w/ label refers to the use of vowel /u/ during dipthong.

reasons for these values is that CzechPD only includes males, and the resulting classification systems can model the speech of these parkinsonian speakers better than in the other corpora in which there are males and females. This suggest that gender-dependent models will provide better results. Likewise, the Czech corpus only includes untreated patients and therefore results suggest that the detection is more efficient in these cases in spite of the fact that most of the patients are in an early stage.

Nevertheless, the main advantage of this work is that it allows to observe the influence of PD in each of the individual phonetic units.

On the other hand, the cross-corpora results suggest that the accuracy of the proposed systems is lower when using a different database for testing than for training, limiting the generalization properties of these systems. However, the sensibility shown in Table 9 is always over 0.87 and the AUC values range from 0.76 and 0.87 depending on the testing corpus. The worst results are obtained when training the models with Neurovoz and CzechPD and testing with GITA. This can be explained by the fact that Neurovoz contains more male speakers than female and CzechPD only includes males. The obtained models are more likely to be adapted to male speakers and GITA is balanced and contains the same number of male and female speakers. Lastly, and summarizing the global results, cross-validation trials (k-folds) provide accuracies between 81% and 94%, with AUC between 0.87 and 0.97 depending on the corpus, while cross-corpora trials provide accuracies between 66% and 76% with AUC between 0.76 and 0.87.

6. Future work and conclusions

In this experimental set several approaches, namely fGMM, GMM-fUBM fGMM-fUBM and opt-fGMM-fUBM are proposed for the automatic detection of PD. The obtained models contain Gaussian densities created or adapted employing only specific phonetic units, allowing to compare independently the features of each phonetic unit between patients and controls for detection purposes.

For the future work in the automatic detection of PD from speech, a comparison of the different segments depending on the manner of articulation or narrowing of the vocal tract must be addressed. In the same manner, and attending to the higher accuracies obtained with CzechPD, the use of gender-dependent models for the detection of PD from speech must be studied. But for these purposes, larger corpora must be employed in order to ensure generalization.

On the other hand, the use of cross-corpora trials is almost nonexistent in PD detection from speech and, thus, its use in this work can be considered a contribution by itself. This type of trials is essential since these demonstrate the generalization capabilities of the proposed approaches. Therefore, this practice must be extended to further works. Regarding the cross-corpora results, compensation techniques to eliminate the effect of the channel must be included to produce more robust models. In the same sense, new score normalization techniques for two-class cases like the studied in this work must be explored.

The accuracies obtained with the proposed approaches suggest that these systems can be used in the clinical practice to support the diagnosis of PD, being part of a multimodal system or in addition to other clinical observations and tests. To this respect, the study of new multimodal systems combining speech with other inputs to support the diagnosis must be addressed.

Additionally, one further step is to clinically test this and other similar systems to prove its usefulness in true clinical environments as there are not published thorough clinical trials of this type.

Lastly, results suggest that PD affects to the articulatory sequence as a whole, influencing more clearly phonetic units requiring a higher narrowing of the vocal tract. For this reason, the analysis of phonetically balanced speech tasks allows to evaluate the presence of PD from speech by using automatic detectors. Additionally, when employing text-dependent utterances as speech tasks, the obtained classification models allow to compare more precisely the acoustic characteristics of the articulation of patients and controls since all the speakers repeat the same sequence of allophones. This is specially relevant in studies and applications containing small corpora for model training and adapting. Consequently, phonetically balanced text-dependent utterances are recommended for automatic detection systems in the clinical practice.

Acknowledgements

The authors of this paper want to thank to Jesus Francisco Vargas Bonilla, Julian David Arias-Londoño and Rafael Orozco-Arroyave from Universidad de Antioquia for sharing the GITA corpus. This work has been supported under the grant of the project *TEC2012-38630-C04-01* and *DPI2017-83405-R* from the Government of Spain, with "Becas de Ayuda a la Movilidad" funded by Universidad Politécnica de Madrid, and with a MIT-Spain Gobal Seed Funds Award.

Appendix

Table 10 includes a list of the International Phonetic Alphabet (IPA) symbols for Castilian Spanish and some examples of the use of the correspondent allophones in Spanish and equivalences in English as well as the associated Kaldi labels in the forced alignment.

Table 10
IPA symbols for the different allophones in Spanish, English equivalent and Kaldi labels.

IPA	Examples	English approximation	Type of sound	Kaldi label
b	balón, vacío, envidia	Best	Voiced bilabial plosive	**b**
ˈβ	bebé, obtener, vivir	Between baby and bevy	Voiced bilabial approximant	**B**
d	dedal, comiendo, aldea	Dead, (putting the tip of the tongue against the upper teeth)	Voiced dental plosive	**d**
ð	dársena, arder, admiración	This	Voiced dental approximant	**D**
f	fase	Face	Voiceless labiodental fricative	**f**
g	gato, lengua, guerra	Got	Voiced velar plosive	**g**
h	Sahara, hall	Hot	Voiceless glottal fricative	**h**
ɣ	trigo, amargo, significado	Go, (without completely blocking airflow on the g)	Voiced velar fricative	**G**
ʝ	ayuno	You	Voiced palatal fricative	**y**
ɟʝ	cónyuge	Job	Voiced palatal affricate	**y**
k	caña, kilo	Scan	Voiceless velar plosive	**k**
l	luz	Lean	Voiced alveolar lateral	**l**
ʎ	llave, pollo	Million	Voiced palatal lateral	**y**

Table 10 (*Continued*)

IPA	Examples	English approximation	Type of sound	Kaldi label
m	madre, campana, anfiteatro	Mother	Voiced bilabial nasal	**m**
n	nido, sin, álbum	Need	Voiced alveolar nasal	**n**
ɲ	España, enyesar	Canyon	Voiced palatal nasal	**ñ**
ŋ	cinco, venga	Sing	Voiced velar nasal	**n**
p	pozo	Spouse	Voiceless bilabial plosive	**p**
r	rumbo, carro, subrayar	*Not present*	Voiced alveolar trill	**R**
ɾ	caro, bravo, partir	Batter (American English)	Voiced alveolar tap	**r**
s	saco, espita, xenón	Sack	Voiceless alveolar fricative	**s**
θ	cereal, encima, zorro	Thing	Voiceless interdental fricative	**s**
t	tamiz	Stand, (putting the tip of the tongue against the upper teeth)	Voiceless dental plosive	**t**
tʃ	chubasco	Choose	Voiceless palatal affricate	**C**
v	afgano	Van	Voiced labiodental fricative	**f**
x	jamón, general, hamster	Scottish loch	Voiceless velar fricative	**j**
z	isla, mismo	Quiz	Voiced alveolar fricative	**s**
a	azahar	Father	Central open vowel	**a**
e	vehemente	Set	Front mid vowel	**e**
i	dimitir, mío, y	See	Front close vowel	**i**
o	boscoso	More	Back mid rounded vowel	**o**
u	cucurucho, dúo	Food	Back close rounded vowel	**u**
j	ligar	Yet	Voiced palatal approximant	**i**
w	cuadro	Wine	Voiced labial-velar approximant	**w**

References

[1] J. Mekyska, Z. Galaz, Z. Mzourek, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, D. Berankova, M. Faundez-Zanuy, et al., Assessing progress of Parkinson's disease using acoustic analysis of phonation, in: Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference on. IEEE, 2015, pp. 111–118.

[2] A. Bandini, F. Giovannelli, S. Orlandi, S. Barbagallo, M. Cincotta, P. Vanni, R. Chiaramonti, A. Borgheresi, G. Zaccara, C. Manfredi, Automatic identification of dysprosody in idiopathic parkinson's disease, Biomed. Signal Process. Control 17 (2015) 47–54.

[3] 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 Dis. (2013) 2013.

[4] F.L. Darley, A.E. Aronson, J.R. Brown, Differential diagnostic patterns of dysarthria, J. Speech Lang. Hear. Res. 12 (2) (1969) 246.

[5] H. Ackermann, W. Ziegler, Articulatory deficits in parkinsonian dysarthria: an acoustic analysis, J. Neurol. Neurosurg. Psychiatry 54 (12) (1991) 1093–1098.

[6] J. Kegl, H. Cohen, H. Poizner, Articulatory consequences of Parkinson's disease: perspectives from two modalities, Brain Cogn. 40 (2) (1999) 355–386.

[7] 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.

[8] A.B. Walsh, Basic parameters of articulatory movements and acoustics in individuals with Parkinson's disease, Mov. Disord. 27 (7) (2012) 843–850.

[9] A. Bandini, S. Orlandi, F. Giovannelli, A. Felici, M. Cincotta, D. Clemente, P. Vanni, G. Zaccara, C. Manfredi, Markerless analysis of articulatory movements in patients with parkinson's disease, J. Voice 30 (6) (2016), 766-e1.

[10] J.A. Logemann, H.B. Fisher, B. Boshes, E.R. Blonsky, Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients, J. Speech Hear. Disord. 43 (1) (1978) 47–57.

[11] J.A. Logemann, H.B. Fisher, Vocal tract control in Parkinson's disease, J. Speech Hear. Disord. 46 (4) (1981) 348.

[12] G. Weismer, M. McNeil, Articulatory characteristics of parkinsonian dysarthria: segmental and phrase-level timing, spirantization, and glottal–supraglottal coordination, Dysarthrias: Physiol. Acoust. Percept. Manage. (1984) 101–130.

[13] J.A. Robbins, J.A. Logemann, H.S. Kirshner, Swallowing and speech production in Parkinson's disease, Ann. Neurol. 19 (3) (1986) 283–287.

[14] E.Q. Wang, L.V. Metman, R.A. Bakay, J. Arzbaecher, B. Bernard, D.M. Corcos, Hemisphere-specific effects of subthalamic nucleus deep brain stimulation on speaking rate and articulatory accuracy of syllable repetitions in Parkinson's disease, J. Med. Speech-Lang. Pathol. 14 (4) (2006) 323.

[15] J. Hlavnivcka, R. Cmejla, T. Tykalova, K. Sonka, E. Ruuzicka, J. Rusz, Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder, Sci. Rep. 7 (1) (2017) 12.

[16] K. Tjaden, J. Lam, G. Wilding, Vowel acoustics in Parkinson's Disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions, JSLHR 56 (5) (2013) 1485–1502.

[17] K. Tjaden, V. Martel-Sauvageau, Consonant acoustics in Parkinson's disease and multiple sclerosis: comparison of clear and loud speaking conditions, Am. J. Speech-Lang. Pathol. 26 (2S) (2017) 569.

[18] 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.

[19] S. Sapir, J.L. Spielman, L.O. Ramig, B.H. Story, C. Fox, Effects of intensive voice treatment (the lee Silverman voice treatment [lsvt]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: acoustic and perceptual findings, J. Speech Lang. Hear. Res. 50 (4) (2007) 899–912.

[20] 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.

[21] N. Roy, S.L. Nissen, C. Dromey, S. Sapir, Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlaryngeal therapy, J. Commun. Disord. 42 (2) (2009) 124–135.

[22] J. Whitfield, A. Goberman, Articulatory acoustic vowel space: application to clear speech in individuals with Parkinson's disease, J. Commun. Disord. (2014).

[23] P. Boersma, Praat: Doing Phonetics by Computer, 2006 http://www.praat.org/.

[24] J. Rusz, J. Hlavnicka, R. Cmejla, E. Ruzicka, Automatic evaluation of speech rhythm instability and acceleration in dysarthrias associated with basal ganglia dysfunction, Front. Bioeng. Biotechnol. 3 (2015).

[25] J.R. Orozco-Arroyave, J. Vasquez-Correa, F. Honig, J.D. Arias-Londono, J. Vargas-Bonilla, S. Skodda, J. Rusz, E. Noth, Towards an automatic monitoring of the neurological state of Parkinson's patients from speech, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP. IEEE (2016) 6490–6494.

[26] J. Godino-Llorente, S. Shattuck-Hufnagel, J. Choi, L. Moro-Velázquez, J. Gómez-Garc’ía, Towards the identification of idiopathic Parkinson's disease from the speech. New articulatory kinetic biomarkers, PLOS ONE 12 (12) (2017), e0189583.

[27] 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.

[28] P.J. Moreno, C. Joerg, J.-M.V. Thong, O. Glickman, A recursive algorithm for the forced alignment of very long audio segments, Fifth International Conference on Spoken Language Processing (1998).

[29] C. Middag, J.-P. Martens, G. Van Nuffelen, M. De Bodt, Automated intelligibility assessment of pathological speech using phonological features, EURASIP J. Adv. Signal Process. 2009 (1) (2009) 629030.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The kaldi speech recognition toolkit, IEEE 2011 workshop on automatic speech recognition and understanding, no. EPFL-CONF-192584. IEEE Signal Processing Society (2011).

[31] aaa, Cepstral analysis technique for automatic speaker verification, IEEE Trans. Acoust. Speech Signal Process. 29 (2) (1981) 254–272.

[32] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted gaussian mixture models, Digit. Signal Process. 10 (1) (2000) 19–41.

[33] J.I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, S. Aguilera-Navarro, P. Gomez-Vilda, An integrated tool for the diagnosis of voice disorders, Med. Eng. Phys. 28 (3) (2006) 276–289.

[34] J. Orozco-Arroyave, J. Arias-Londoño, 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).

[35] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mariño, C. Nadeu, Albayzín speech database: Design of the phonetic corpus, Eurospeech 1993. Proceedings of the 3rd European Conference on Speech Communication and Technology, 1. ISCA (1993) 175–178.

[36] L. Moro-Velazquez, J.A. Gomez-Garcia, J.I. Godino-Llorente, J. Villalba, J.R. Orozco-Arroyave, N. Dehak, Analysis of speaker recognition methodologies

and the influence of kinetic changes to automatically detect Parkinson's disease, Appl. Soft Comput. 62 (2018) 649–666.
[37] 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.
[38] 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.
[39] J.R. Duffy, Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, Elsevier Health Sciences, 2013.
[40] 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.
[41] 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.
[42] A.J. Hughes, S.E. Daniel, L. Kilford, A.J. Lees, Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases, J. Neurol. Neurosurg. Psychiatry 55 (3) (1992) 181–184.
[43] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, G. Logroscino, Accuracy of clinical diagnosis of parkinson disease a systematic review and meta-analysis, Neurology 86 (6) (2016) 566–576.
[44] A. Lee, R.M. Gilbert, Epidemiology of parkinson disease, Neurol. Clin. 34 (4) (2016) 955–965.
[45] Y. Lei, N. Scheffer, L. Ferrer, M. McLaren, A novel scheme for speaker recognition using a phonetically-aware deep neural network, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP, IEEE (2014) 1695–1699.