Automatic detection of Parkinson's disease from components of modulators in speech signals

Detección automática de la enfermedad de Parkinson usando componentes moduladoras de señales de voz

DOI: https://doi.org/10.17981/cesta.01.01.2020.05

Artículo de investigación científica. Fecha de recepción: 09/09/2020 Fecha de aceptación: 29/09/2020

Jhon F. Moofarry
Universidad Santiago de Cali. Cali (Colombia)
jhon.moofarry00@usc.edu.co

Patricia Argüello-Velez
Universidad Santiago de Cali. Cali (Colombia)
patricia.arguello00@usc.edu.co

Milton Sarria-Paja
Universidad Santiago de Cali. Cali (Colombia)
milton.sarria00@usc.edu.co

Para citar este artículo:
J. Moofarry, P. Argüello-Velez & J. Sarria-Paja, "Automatic detection of Parkinson's disease from components of modulators in speech signals", *J. Comput. Electron. Sci.: Theory Appl.*, vol. 1, no. 1, pp. 71–82, 2020. https://doi.org/10.17981/cesta.01.01.2020.05

Abstract— Parkinson's Disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease. This disorder mainly affects older adults at a rate of about 2%, and about 89% of people diagnosed with PD also develop speech disorders. This has led scientific community to research information embedded in speech signal from Parkinson's patients, which has allowed not only a diagnosis of the pathology but also a follow-up of its evolution. In recent years, a large number of studies have focused on the automatic detection of pathologies related to the voice, in order to make objective evaluations of the voice in a non-invasive manner. In cases where the pathology primarily affects the vibratory patterns of vocal folds such as Parkinson's, the analyses typically performed are sustained over vowel pronunciations. In this article, it is proposed to use information from slow and rapid variations in speech signals, also known as modulating components, combined with an effective dimensionality reduction approach that will be used as input to the classification system. The proposed approach achieves classification rates higher than 88 %, surpassing the classical approach based on Mel Cepstrals Coefficients (MFCC). The results show that the information extracted from slow varying components is highly discriminative for the task at hand, and could support assisted diagnosis systems for PD.

Keywords— Modulation spectrum; Parkinson's disease; speech signals; pattern recognition; covariance features

Resumen— La Enfermedad de Parkinson (EP) es el segundo trastorno neurodegenerativo más común después de la enfermedad de Alzheimer. Este trastorno afecta principalmente a los adultos mayores con una tasa de aproximadamente el 2%, y aproximadamente el 89% de las personas diagnosticadas con EP también desarrollan trastornos del habla. Esto ha llevado a la comunidad científica a investigar información embebida en las señales de voz de pacientes diagnosticados con la EP, lo que ha permitido no solo un diagnóstico de la patología sino también un seguimiento de su evolución. En los últimos años, una gran cantidad de estudios se han centrado en la detección automática de patologías relacionadas con la voz, a fin de realizar evaluaciones objetivas de manera no invasiva. En los casos en que la patología afecta principalmente los patrones vibratorios de las cuerdas vocales como el Parkinson, los análisis que se realizan típicamente sobre grabaciones de vocales sostenidas. En este artículo, se propone utilizar información de componentes con variación lenta de las señales de voz, también conocidas como componentes de modulación, combinadas con un enfoque efectivo de reducción de dimensiónalidad que se utilizará como entrada al sistema de clasificación. El enfoque propuesto logra tasas de clasificación superiores al 88 %, superando el enfoque clásico basado en los Coeficientes Cepstrales de Mel (MFCC). Los resultados muestran que la información extraída de componentes que varían lentamente es altamente discriminatoria para el problema abordado y podría apoyar los sistemas de diagnóstico asistido para EP.

Palabras clave— Espectro de modulación; Enfermedad de Parkinson; señales de voz; reconocimiento de patrones; características de covarianza

I. Introduction

PARKINSON'S Disease (PD) is a progressive neurodegenerative disorder that mainly affects motor system. The loss of dopamine-containing neurons in the midbrain is progressive and affects different parts of the nigral complex to different degrees [1]. Given the control loss of motor activities, 90 % of PD patients develop different speech disorders, being phonation related problems the first to manifest [2]. Parkinson's disease also present signs such as slowness, tremor, stiffness, and postural instability, as for speech, recent research suggest that PD affects different speech production dimensions, such as breathing, phonation, articulation and prosody [3], [4] which reflects in signs such as reduced intensity, monotonous and rough speech, with imprecise articulation, and lack of fluidity [5].

Findings from speech and voice pathologies research show that disorders affecting motor skills, such as PD, manifest in dysarthric speech, which is characterized by alterations not only in the excitation source (air from the lungs and vocal fold vibrations), but also altering the syllabic rate, the general temporal dynamics characteristics of the generated speech signal, and intelligibility [6]-[8]. Advances in signal processing and machine learning techniques allow to implement data driven applications to assist speech pathologist not only at early stages but also during advanced stages of the disease, and help the decision making process [9]-[12].

As some examples, can mention researchs [13], [14] where it was explored noise, periodicity and stability measures, as well as modulation spectral based features, some results show an accuracy of 71 % for the sustained vowel */i/* using modulation spectral features. However, best results were reported when using stability and periodicity features, for vowel */a/* with an accuracy of 91 %. More recently, it was reported that unvoiced segments contain highly discriminative information [15]. According to results reported on cross-language experiments, classification rates between 85 % to 99 % depending on the language and the speech task, could be attained. One important drawback of this approach is the requirement of a precise measurement of Voice Onset Time (VOT), which is difficult to do using an automatic algorithm.

From the linguistic and clinical point of view, speech signals analysis requires clarity in the linguistic structure of the stimuli, taking into account phonetic balance and changes in the recordings conditions according to the specific needs of the intended study. The Diadochokinetic tasks (DDK) are characterized by a direct syllabic construction, i.e., consonant (C) +vowel (V), in an alternate way during an expiratory breath. These linguistic structures allow to explore information during the phonation of unvoiced consonants followed by a voiced sound (vowel) from a linguistic and clinical point of view [16], [17]. Information from these analyses allows to study supraglottic articulatory phenomena, associated with the unvoiced segment, whilst the voiced segment allows to study phonatory phenomena, specifically glottal phenomena [18].

Exploring DDK with the alternation */pataka/* is widely known as a linguist stimuli for indirect instrumental studies aimed at the diagnosis of motor speech disorders [19]. The consonants */p/*, */t/* and */k/* are part of a class of sounds called stops or plosives, unlike other sounds, which can be described largely in terms of steady-state spectra, stops are transient acoustically complex phonemes, with different acoustic aspects depending on where the closure occurs. For example, */k/* requires the occlusion of the back of the tongue against the soft palate, */t/* requires closure in the vocal tract (tongue against the dental alveoli), and */p/* at the lips. With the chest muscles continuing to attempt to expel air, pressure builds behind the closure until it is released by opening the occlusion [20]. The temporal transition from the consonant to the vowel contains acoustic information related to the articulatory precision and quality of laryngeal coordination to start the vowel production [21], [22].

For this study, were considered recordings with DDK in sequences containing a plosive unvoiced sound followed by a vowel. Considering that the order in which the DDK task is articulated can provide important information for diagnosis purposes and it is proposed to explore the slow varying envelopes of the speech signal to extract such information. You have the hypothesize that information from slow varying envelopes of the speech signal can be associated to the relative slow transition from one point of articulation to the other, and this can give insights related to the coordination or skill level the speaker has to change the vocal tract configuration to utter a given sequence. This will be explored using previously proposed acoustic features that explore information from modulation components.

The remainder of this paper is organized as follows. Section 2 provides a theoretical background from an articulatory point of view as well as the machine learning and signal processing techniques to be used. Section 3 describes the corpus employed for the experiments, the validation strategy and the settings for the classification system. Section 4, presents the results and analysis of our experiments and the performance achieved by the proposed schemes. Lastly, Section 5 presents the final conclusions.

II. Theoretical Background

A. *Diadochokinetic tasks - (DDK)*

Human speech is a natural and flexible mode of communication that not only serves to communicate information from a speaker to one or more listeners, it also conveys traits such as identity, age, gender, social and region of origin, emotional, and health states, to name a few [20], [23]. The speaker produces a speech signal in the form of pressure waves, air from the lungs causes the vocal folds to vibrate, exciting the resonances of the vocal tract in a particular configuration. This configuration modulates the excitation source allowing the speaker to produce a great variety of voiced sounds [20]. Speech is produced from a time varying vocal tract system, which makes speech signals dynamic or time-varying in nature. Even though the speaker has control over many aspects of speech production, e.g., loudness, voicing, or vocal tract configuration, much speech variation is not under speaker control and is random, e.g., vocal fold vibration is not truly periodic.

From the linguistic and clinical perspective the DDK sequences */pataka/* and */pakata/* explore the place of articulation in the anteroposterior (from front to back) sense, i.e., */p/* labial, */t/* teeth and tongue and */k/* velar, keeping concordance with the contact area of articulators. The DDK */pataka/* results effective given such articulatory distribution, allowing a muscular order when uttering the sequence (Fig. 1a). On the other hand, the DDK */pakata/*, does not keep this order in the movement of articulators, as it does a transition from lips to the soft palate (velum) and from there, to the anterior teeth (Fig. 1b). Thus, articulation time for such construction is longer as it requires more coordination to achieve articulatory precision.

From the acoustic point of view, by analyzing the burst in plosive consonants the place of articulation can be characterized by extracting the frequency where there is a maximum energy concentration. For Spanish, we have that */pa/* has a maximum energy concentration between 500 and 1000 Hz, for */ta/* between 2000 and 3500 Hz, and for */ka/* between 1000 and 2000 Hz [25]. Having these ranges, we illustrate the spectral characteristics of the two DDK sequences in Fig. 2.

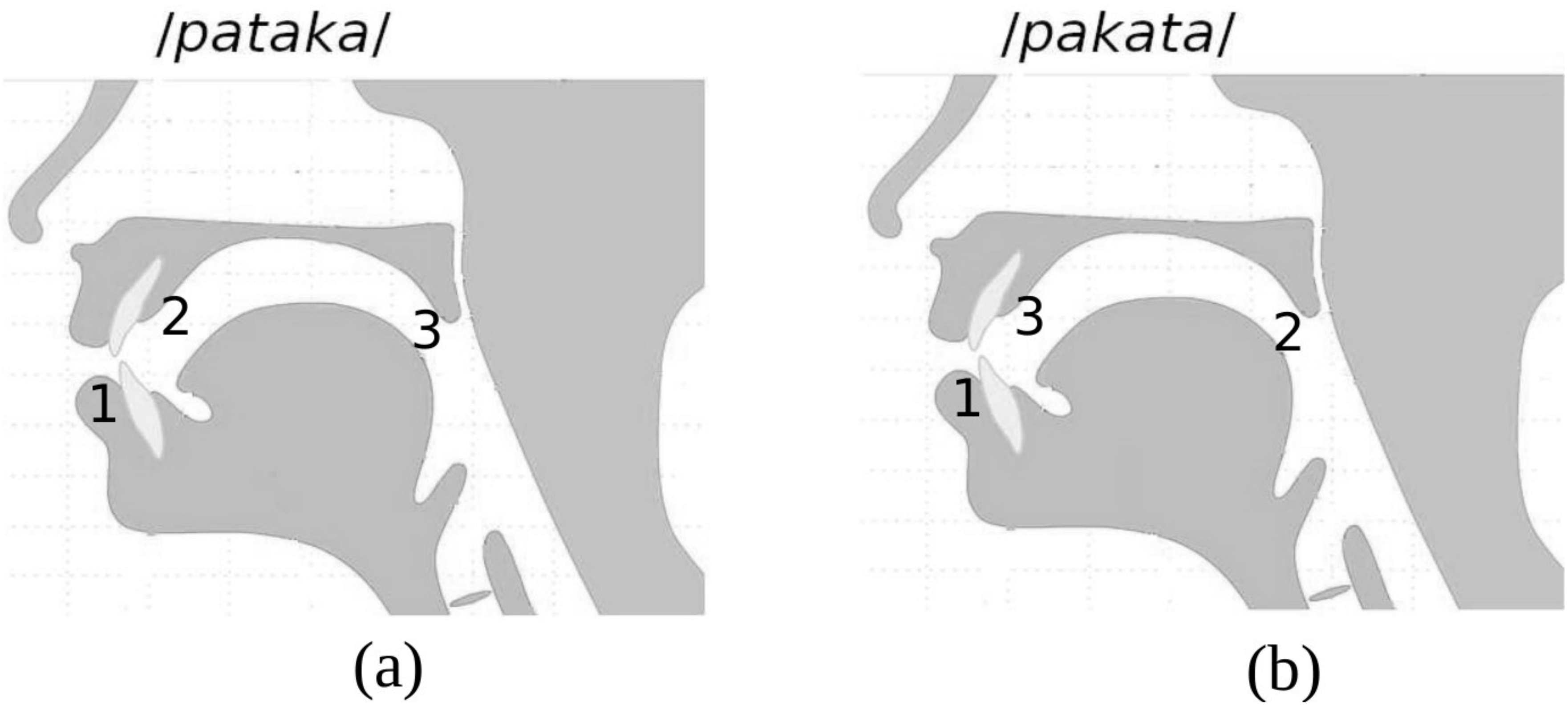

Fig. 1. Place of articulation order for: (a) /pataka/ lips (1) teeth and tongue (2) and velum (3). (b) /pakata/, lips (1) velum (2) and teeth and tongue (3).
Source: Adapted from [24].

B. *Speech characterization*

Spectral or frequency analysis methods applied over “short time” duration frames, have been the preferred approaches to extract information from speech signals for many applications [23], [26], [20], [27] and are the basis for any speech enabled application. In this section describe in more detail some of the typical feature extraction methods that will be explored within this research.

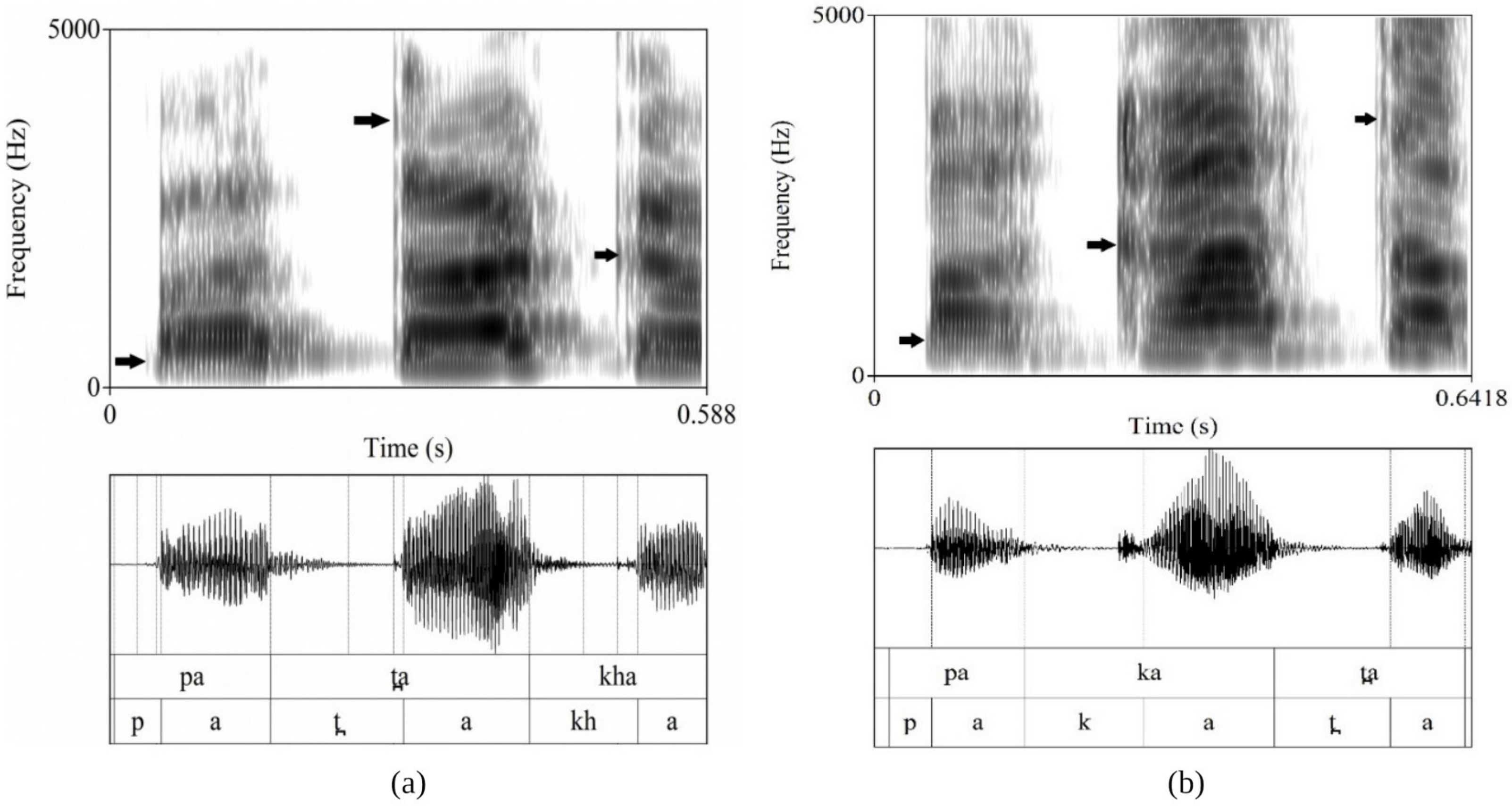

Fig. 2. Spectrograms comparing the two DDK sequences (a) /pataka/ and (b) /pakata/. The arrows indicate the maximum energy concentration for the consonants /p/, /t/ and /k/.
Source: The authors.

1) *Mel Frequency Cepstral Coefficients*

The most popular analysis method for automatic speech recognition combines cepstral analysis theory [28] with aspects related to the human auditory system [20]. The so-called mel-frequency cepstral coefficients (MFCC) are the classical frame based feature extraction method widely used in speech applications. One of the reasons for widespread usage of MFCCs is that they provide an alternative and efficient representation for speech spectra which incorporates some aspects of audition [20], [29].

For MFCC computation, each speech recording is pre-emphasized and windowed in overlapped frames of length τ using a Hamming window to smooth the discontinuities at the edges of the segmented speech frame. To allow $x(n)$ represent a frame of speech that is pre-emphasized and Hamming-windowed. First, $x(n)$ is converted to the frequency domain by an N point Discrete Fourier Transform (DFT) and the resulting energy spectrum can be written as $|X(k)|^2$, with $1 \leq k \leq N$. Next, P triangular band pass filters spaced according to the mel scale are imposed on the spectrum. These filters do not filter time domain signals, they instead apply a weighted sum across the frequency indexes k, which allows to group the energy of frequency bands into a single value, resulting in P energy values $E(l)$ with $1 \leq l \leq P$. Finally, a Discrete Cosine Transform (DCT) is applied to the log-filterbank energies.

The temporal changes in adjacent frames play a significant role in human perception. To capture this dynamic information in the speech, first- and second-order difference features (Δ and ΔΔ MFCC) can be appended to the static MFCC feature vector [20], [29].

2) *AM-FM based features*

Different types of low-level features have been proposed for speech processing and representation aiming to improve the performance of MFCC baseline systems under noisy/reverberant conditions, or to provide complementary information to MFCCs [30]. Some of these features are extracted from slowly varying subband envelopes, which intend to explore as an alternative to previously proposed features for PD diagnosis. As an example, features derived from the AM-FM signal representation [31] have proven to be more robust in noisy conditions and perform at the same level as cepstral coefficients in clean conditions [32], [33]. The main difference is that cepstral coefficients are based on power spectrum estimation (i.e., frequency domain) whilst features derived from the AM-FM signal representation are computed in the time domain. More specifically, the AM-FM model decomposes the speech signal into bandpass channels and characterizes each channel in terms of its envelope and phase (instantaneous frequency) [32], [34]. The speech signal s(n) is filtered through a bank of N_K filters, resulting in the bandpass signal $y_k(n) = s(n) * h_k(n)$, where $h_k(n)$ corresponds to the impulse response of the k-th filter (* denotes convolution). After filtering, each analytic sub-band signal $s_k(n)$ is uniquely related to a real valued bandpass signal $y_k(n)$ by the relation (1).

$$\mathrm{s_k(n) = y_k(n) \ast j\ y^k(n)} \tag{1}$$

Where $y_k(n)$ stands for Hilbert transform of $y_k(n)$. Here, two features are explored based on the AM-FM signal decomposition. The first is the so called Weighted Instantaneous Frequencies (WIF). These features are computed by combining the values of a low–frequency modulator denoted as $m_k(n)$ and the instantaneous frequency, denoted as $f_k(n)$, per bandpass signal using a short-time approach [32] in the following way (2):

$$F_k = \frac{\sum f_k(i) m_k^2(i)}{\sum m_k^2(i)}, \ k = 1, \ldots N_k \tag{2}$$

Fk is calculated over the full length of the signal with increments of *τ/2*. The second feature set is the Mean Hilbert Envelope Coefficients (MHEC) [33]. In this case, the envelope $m_k(n)$ is blocked into frames and the mean Hilbert envelope for a specific frame in the channel *k* is calculated (3):

$$E_k = \frac{\log\left(\frac{1}{\tau}\sum \omega(i - n_0 + 1) m_k(i)\right)}{\bar{E}_k}, \ k = 1, \ldots N_k \tag{3}$$

Where ω(*n*) is a Hamming window of length τ, and the term represents the long-term average in each channel which normalizes the values of E_k. Finally, for a specific frame and using all Ek values, a DCT is applied to produce the MHEC features [33].

3) *Amplitude modulation features*

Modulation spectrum based features have been explored in the past for different purposes, such as neutral/whispered speech classification [35], or speech and speaker recognition in reverberant environments [36], [37], thus alleviating effects of acoustic environment in processing of speech. In the speaker recognition field, the modulation frequency (modulation domain) represents the frequency content of the subband amplitude envelopes and conveys information about speaking rate and other speaker specific attributes [38], hence this is an alternative way to the AM-FM model when exploring slowly varying subband information. In [39] were proposed the Auditory-inspired Amplitude Modulation Features (AAMF), using time contexts or blocks of spectrograms consisting of multiple consecutive short-time frames. In particular, each recording is represented as a tensor with dimensions corresponding to acoustic frequency, modulation frequency and time. Hence, each time context is represented by a $N_{fa} \times N_{fm}$ matrix (Number of acoustic bands × Number of modulation bands). This representation can be collapsed into a vector, followed by *log*10 compression, then each feature vector is projected to a lower dimensional space using Principal Component Analysis (PCA) [39].

This approach, together with previously described short-time approaches, can be explored when using models to capture information directly from time series, such as Hidden Markov Models (HMM), however, will also explore to use the covariance matrix features before applying PCA as described below in order to use these frame based features in a standard classification system.

C. *Dimensionality reduction*

This stage relies on a suitable change (simplification or enrichment) of a representation, e.g. by a reduction of the number of features, relations or primitives describing objects, or some non-linear transformation of the features, to enhance the class or cluster descriptions [40]. For instance, if assume the output of the triangular filterbanks in the MFCC pipeline to be the features, then the DCT can be seen as a feature transformation process, which reduces the dimensionality and decorrelates the variables, thus resulting in a more compact and informative feature vector. Techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) are commonly used for more general classification tasks [40].

Herein, explore an approach that has been proposed before for speaker recognition applied to speakers with dysarthria [41]. Such an approach is based on the covariance matrix. Covariance Features (CF) are a simple representation of data using the sample covariance matrix. This type of feature representation wasfirst used as a region descriptor for the problems of object detection and texture classification from images [42], and extended to video [43] and audio [44] processing. For a given input speech signal first extract a sequence of short-time features. The sample covariance matrix is then computed for such a sequence and the covariance feature vector is obtained by collapsing into a vector the upper (or lower) triangular part of the covariance matrix and used as standard features [41]. Finally, applies PCA to reduce the dimensionality of the resulting feature vector, retaining 98% of cumulative variance which results in feature vectors of about 40 to 50 components, depending on the short-time feature representation.

D. *Classification - generalization and inference*

Once a set of features or parameters to describe the speech recordings, another important stage is the generalization/inference stage. In this stage, a classifier/identifier is trained. The training process involves the parameter tuning of models to describe training samples, i.e., features extracted from speech recordings. The learning process requires assumption son the general form of the model or the classifier, and uses the training samples to estimate the unknown parameters of the model. Then an algorithm is applied in order to reduce the error on a set of training data or in general terms, optimize a cost function related to the task at hand [45]. First, will use a model to describe the dynamic information embedded in short time or frame based representations of speech signals. For this purpose, and due to the simplicity and efficiency of its parameter estimation algorithm, the Hidden Markov Model (HMM) was for many years the dominant approach for modeling discrete time series, finding widespread application in the areas of speech processing [46]. For this reason used a HMM based classification system as our baseline approach.

1) *Hidden MarkovModels-HMM*

Are a general statistical modeling technique for sequences or time series. The HMM is composed of a number of states (n_ϑ), each state emits symbols (observations) according to symbol-emission probabilities, and the states are interconnected by state-transition probabilities. Starting from some initial state, a sequence of states is generated by moving from state to state according to the state-transition probabilities until an end state is reached. Each state then emits symbols according to the state's emission probability distribution, creating an observable sequence of symbols [46].

The model parameters, denoted as $\lambda = (A, B, \pi)$, include: (i) an initial state $\pi = [p_1, ..., p_{n\vartheta}]^T$ with n elements, $n \in [1, n_\vartheta]$, describing the distribution over the initial state set, (ii) a transition matrix $A \in R\ n_\vartheta \times n_\vartheta$ with elements a_{ij}, $i, j \in [1, n_\vartheta]$ to denote the transition probability to node j, given that the HMM is currently in state i; and (iii) an observation matrix $B = \{b_j(\cdot)\}$ that represents the observation distribution per state j in the model. It employs parametric distributions of a predetermined form that mostly are based on weighted sums (mixtures) of multivariate Gaussian densities [46].

2) *Support Vector Machines – SVM*

The SVM is a binary classifier which models the decision boundary between two classes as a separating hyperplane. In our experiments, two classes are involved: a "positive class", i.e., individuals with PD and a "negative class", i.e., individuals from the control group (HCs). By using labelled training vectors, the SVM optimizer finds a separating hyperplane that maximizes the separation between these two classes [47], [48]. The discriminant function is given by (4):

$$f(x) = \sum_{j=1}^{N} a_j c_j K(x, x_j) + b \tag{4}$$

Where $c_j \in \{+1, -1\}$, are the labels for the training vectors. The kernel function $K(\cdot,\cdot)$ is constrained to have certain properties (the Mercer condition). The support vectors x_j, their weights a_j and the bias term b, are determined during training [47], [48].

3) *K-Nearest Neighbors – KNN*

In the K-nearest neighbor classifier it is wanted to minimize the probability of misclassification. Thus, to classify a new point, are identified the K-nearest points from the training data set and then assign the new point x to the class having the largest number of representatives amongst this set. The particular case of $K = 1$ is called the nearest-neighbor rule, because a test point is simply assigned to the same class as the nearest point from the training set [40].

III. Experimental Setup

A. *Speech stimuli*

An evaluation corpus containing speech recordings from 100 participants sampled at 44.1 KHz with 16 resolution-bits was used. This speech corpus contains recordings of 50 patients with PD and 50 healthy individuals (control group - HCs). These recordings were captured in noise controlled conditions, in a sound proof booth. The participants are balanced by gender and age (Fig. 3). The participants with a positive diagnosis for PD were diagnosed and labeled by neurologist experts. The labels of their neurological evaluation were assigned according to the UPDRS-III and Hoehn & Yahrscales [14].

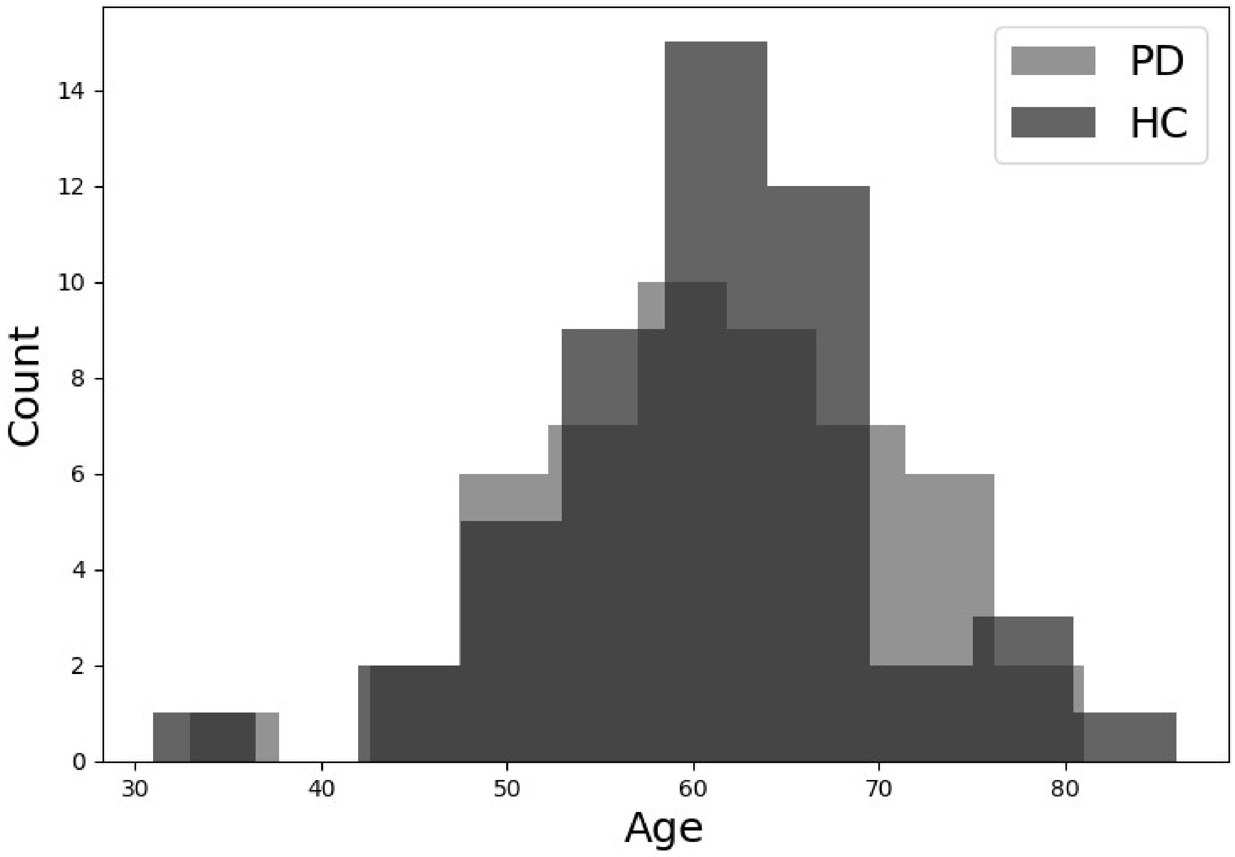

Fig. 3. Histogram showing age vs count of participants in the database.
Source: The authors.

For the experiments herein, we used recordings containing successive repetitions of the utterances /*pataka*/ and /*pakata*/. As was mentioned before, these kind of linguistic constructions allow to explore information related to the articulatory precision and quality of laryngeal coordination, as well as level of motor coordination of all other muscles and organs that can be affected by PD [26]. Analyzing speech signals from DDK will also allow us to explore the relationship of information extracted from slowly varying subband envelopes to PD.

B. *Feature Extraction*

Prior to apply any feature extraction process, eachspeech recording was downsampled to 16 kHz. Next the recordings were pre-emphasized using a first order finite impulse response filter with constant a= 0.97. Features such as MFCC, WIFs and MHEC, were computed on a per-window basis using a 25 ms window with 15 ms overlap. For MFCC computation, 27 triangular bandpass filters spaced according to the mel scale were used to compute 13 MFCC features including the 0–th order cepstral coefficient (log-energy). Dynamic or transitional features are computed bymeans of an anti-symmetric Finite Impulse Response (FIR) filter with nine coefficients to avoid phase distortion of the temporal sequence.

For WIFs and MHEC, a gammatone filterbank [49] with 27 channels was used. Filter center frequencies range from 1000 Hz to 7000 Hz and their bandwidths are characterized by the mel frequency scale. For the AAMF features, time contexts are 200 ms long, frame length and overlap is adjusted to guarantee a 80 hz bandwidth in the modulation domain. The number of filters in the acoustic domain is set to N_{fa}= 27 filters, with filter center frequencies distributed according to the mel scale, and in the modulation domain to N_{fm} = 8 filters, using logarithmically-spaced triangular bandpass filters distributed between 0.01–80Hz.

C. *Classification system and validation strategy*

For the experiments presented here, the baseline has been adjusted using an HMM-based classifie, and the best adjustment will be chosen by varying the number of states and the number of Gaussians per state, this will be done using as input the classical MFCC feature vectors. After, setting the baseline, it is compared with the other feature extraction methods such as WIFs, MHEC and AAMF.

Next, by using the approach described in Section 2.3, it was possible to map the variable length frame based representation to a fixed dimensional feature vector. This allows the use of classic classification approaches such as GMM, SVM and KNN. The hyper-parameters for each classification system are tuned by using a grid search and selecting those parameters that present the best average accuracy.

Finally, random cross-validation (10-fold) was used with 80% of the input data recordings kept for system training and 30% left for validation [45]. Classification accuracy is reported as average accuracy ± the standard deviation across the 10 cross-validation trials. It is important to note that the folds are randomly assembled with the constraint of the balance of age and gender of the speakers as suggested by [15].

IV. Experimental results and discussion

A. *Baseline system characterization*

Table I and Table II report the accuracy results for classification using a HMM based classification system. In order to establish a baseline performance, the HMM was first tuned using only the MFCC feature extraction algorithm. This is achieved by varying the number of states and number of Gaussians per state. The first important result to be highlight is the high variability in all accuracy values across the 10 cross-validation trials, this is independent of the DDK sequence, the number of states or the number of Gaussians per state. These results suggest that HMM might not be the best choice to model the information encoded in speech signals when trying to predict PD. All results, nevertheless, are better than chance (50%), thus showing that there is discriminative information in the classical MFCC features for the task at hand. Also it is observed that for /pataka/ a small number of Guassians and independent of the number of states seems to work best, whilst for /pakata/ highest accuracy results were achieved with a small number of states, this however is not conclusive given the high variability of the results.

Table I.
Accuracy (%) results for the sequence /pataka/ with HMM varying the number of states and the number of Gaussians per state.

No. of States	No. of Gaussians			
	2	4	8	16
2	75.5±7.2	74.0±8.4	72.5±10.6	74.0±9.9
3	74.0±8.7	70.0±11.0	73.0±10.3	69.0±10.4
5	77.5±7.1	75.0±7.8	71.5±10.5	69.5±10.3

Source: Authors.

Table II.
Accuracy (%) results for the sequence /pakata/ with HMM varying the number of states and the number of Gaussians per state.

No. of States	No. of Gaussians			
	2	4	8	16
2	76.0±7.3	74.5±12.1	74.0±5.5	76.0±7.0
3	70.0±8.1	75.5±6.8	71.5±6.6	72.5±4.2
5	75.5±7.9	75.5±6.8	73.5±8.1	73.5±6.6

Source: Authors.

Next, is then compared the MFCC with other feature extraction techniques, results are reported in Table III. As can be seen, despite the high variability of HMM+MFCC, this strategy performs better than features derived from the AM-FM model, or the modulation spectrum approach. Something important to highlight, is the slight difference, for all feature extraction methods, in favor of the sequence /pataka/. As mentioned before, this sequence respects an articulatory order: from front to back.

Table III.
Accuracy(%) results for the sequences /pakata/ and /pataka/ with HMM varying the feature set. For /pataka/ we used five states and two Gaussians per state. For /pakata/ we used two states and two Gaussians per state.

Features	/pakata/	/pataka/
MFCC	76.0±7.3	77.5±7.1
WIF	63.5±8.1	64.5±8.3
MHEC	60.0±4.1	67.5±8.5
AAMF	66.5±8.1	73.0±7.8

Source: Authors.

B. *Feature mapping using covariance matrix features*

After setting the baseline using the HMM based classifier, we evaluate a different strategy. The approach is considered not to be in the machine learning algorithm, it should be instead at the feature extraction process and the preprocessing prior to feeding feature vectors to a classifier. At this stage the approach described above (Section 2.3) tomap a variable length frame based representation to a fixed dimensional feature vector. Using this approach you can to implement classical pattern classification strategies such as SVM or KNN, and compare these two strategies against the Naive Bayes Classifier (NBC).

Results are presented in Table IV and Table V, presenting a comparison for all feature sets and three different classifiers. Furthermore, the results of the two DDK tasks, i.e., */pakata/* (Table IV) and */pataka/* (Table V), show two important results when comparing with previous section. First, the performance of the baseline feature set, MFCC, is not affected when using SVM combined with the feature extraction method from Section 2.3, in fact, we observe similar performance for WIF and MHEC, with MHEC having an improvement of around 10 % when using the sequence */pakata/*. Finally, the AAMF feature set attains the better improvements, as it goes from 66.5 % accuracy to 85.1 % with the sequence */pakata/*, and from 73.0 % accuracy to 88.0 % with the sequence */pataka/*. These results show that the feature extraction strategy is effective, as it maintains highly discriminative information in all feature sets and helps to reduce the computational burden when compared with HMM. While with HMM it is required around 20 seconds (average) per feature set to train a model, training a SVM based classifier takes around 0.9 seconds (average).

Table IV.
Accuracy (%) results for the sequence /pakata/ comparing the feature sets by using different classification approaches.

Feature set	/pakata/		
	NBC	KNN	SVM
MFCC	76.1±3.0	70.0±2.5	78.0±3.2
WIF	60.2±3.0	60.0±5.1	62.2±4.0
MHEC	61.4±4.0	76.1±2.5	64.3±3.7
AAMF	69.0±5.0	80.3±3.1	85.1±2.0

Source: Authors.

Table V.
Accuracy (%) results for the sequence /pataka/ comparing the feature sets by using different classification approaches.

Feature set	/pataka/		
	NBC	KNN	SVM
MFCC	57.1±2.0	68.1±2.0	77.1±2.0
WIF	54.3±2.9	65.2±4.2	66.1±4.0
MHEC	55.0±3.2	75.1±2.5	78.3±3.1
AAMF	70.1±4.0	71.0±3.1	88.0±2.8

Source: Authors.

The second important result, except for MFCC, is that all systems present a better performance when using the DDK sequence */pataka/* (Table V). This is important for the WIF, MHEC and AAMF feature sets, as all three sets in some way are looking into information encoded in slow varying envelopes. This observation allows to have insights related to the relevance of the phonetic structure used in the DDK sequence */pataka/*, and its articulatory distribution anteroposterior (lips, teeth, velum). In [50], authors presented the idea that the movements at the lips for */p/* and the lingual apex for */t/* are faster when comparing to the movement at the posterodorsal region (close to the velum) for */k/*, which implies a longer transition towards a bigger contact area. Taking this into account, the sequence */pataka/* is a linguistic task that follows an articulatory order from a physiological point of view. And from an acoustic point of view, if we measure the voice onset time (VOT) for these unvoiced stop consonants, we confirm a increased distribution in the following order: */p/* < */t/* < */k/*.

V. Discussion and Conclusions

This paper has addressed the issue of Parkinson's Disease (PD) detection based on information extracted from speech signals. Two different diadochokinetic tasks where used in order to explore information during the phonation of unvoiced consonants followed by a voiced sound. Two approaches were explore to implement the classification system, first a HMM based system to model the dynamic of variable length representations as is the case of frame based features for speech signals. Second, a feature extraction method was used to map these variable length frame-based representations to a fixed dimensional feature space in order to use typical classification strategies.

As PD affects motor skills, it is expected to observe some symptoms in speech production. The main hypothesis explored in this paper is that as coordination of articulator is affected, then long term slow varying information encoded in speech can signal and be linked to some disorders affecting motor skills as is the case of PD, that opposed to previous research in the field, where information from stability and periodicity computed in a short term basis has been preferred for the task at hand. In this regard, explored three feature sets that extract information from slow varying envelopes in speech signals, i.e., WIF, MHEC and AAMF, features that have shown to be highly informative in other speech enabled applications. As a result, found that in general these features perform poorly when compared to the classical MFCC + HMM paradignm. However, in a different scenario, where the feature set is mapped to a static feature space, the AAMF feature set shows to be highly informative. There are absolute differences of around 10 % in accuracy when comparing with MFCC. These results are part of an ongoing research, looking at the discriminative information extracted from modulation spectral representation of speech signals. These results show that AAMF features contain discriminative information and need to be explored in detail for the task at hand.

Furthermore, found that selecting the DDK task actually has some influence in the final results. As was observed, in the sequence /pataka/ is preferred for the task at hand as its efectivenes is supported from a clinical and linguistic point of view that has been previously established, given the articulatory distribution, i.e., an anterioposterior order for the articulators and the increment of the VOT for the stop consonants */p/*, */t/* and */k/* as the sequence advances [51]. This motivates us to continue studying the spectral and acoustic features of these DDK sequences as well the unvoiced stop consonants, and vocal sounds that compose them, in order to find more insights related to elements that affect or have influence in the envelope of the phonetic sequence for the diagnosis of Parkinson's disease.

Acknowledgment

The authors acknowledge funding from Univesidad Santiago de Cali (Colombia).

References

[1] J. M. Fearnley & A. J. Lees, "Ageing and parkinson's disease: substantia nigra regional selectivity", *Brain*, vol. 114, no. 5, pp. 2283–2301, Oct. 1991. https://doi.org/10.1093/brain/114.5.2283

[2] P. Gómez-Vilda, D. Palacios-Alonso, V. Rodellar-Biarge, A. Álvarez-Marquina, V. Nieto-Lluis & R. Martínez-Olalla, "Parkinson's disease monitoring by biomechanical instability of phonation", *Neurocomputing*, vol. 255, pp. 3–16, Sept. 2017. https://doi.org/10.1016/j.neucom.2016.06.092

[3] T. Khan, J. Westin & M. Dougherty, "Classification of speech intelligibility in parkinson's disease", *BBE*, vol.34, no. 1, pp. 35–45, Jan. 2014. https://doi.org/10.1016/j.bbe.2013.10.003

[4] J. Rusz, R. Cmejla, H. Ruzickova & E. Ruzicka, "Objectification of dysarthria in parkinson's disease using bayes theorem", in Proc. *10th NEHIPISIC*, WSEAS, CGK, ID, Dec. 1-3, 2011, pp. 165–169. https://dl.acm.org/doi/10.5555/1959586.1959620

[5] L. O. Ramig, C. Fox & S. Sapir, "Speech treatment for parkinson's disease", *Expert Rev Neurother*, vol. 8, no. 2, pp. 297–309, Feb. 2008. https://doi.org/10.1586/14737175.8.2.297

[6] P. C. Doyle, H. A. Leeper, A.-L. Kotler, N. Thomas-Stonell, C. O'Neill, M.-C. Dylke & K. Rolls, "Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility", *JRRD*, vol. 34, no. 3, pp. 309–316, Jul. 1997. Available: https://www.rehab.research.va.gov/jour/97/34/3/pdf/doyle.pdf

[7] J. R. Duffy, *Motor speech disorders e-book: Substrates, differential diagnosis, and management*, St. Louis, Mo, USA: Elsevier Health Sciences, 2013.

[8] 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*, vol. 32, no. 3, pp. 141–186, May. 1999. https://doi.org/10.1016/s0021-9924(99)00004-0

[9] T. H. Falk, W.-Y. Chan & F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility", *Speech Commun*, vol. 54, no. 5, pp. 622–631, Jun. 2012. https://doi.org/10.1016/j.specom.2011.03.007

[10] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin & S. Frame, "Dysarthric speech database for universal access research", in *INTERSPEECH 2008*, ISCA, BRN, AUS, Sep. 22-26, 2008. Available at: https://www.isca-speech.org/archive/archive_papers/interspeech_2008/i08_1741.pdf

[11] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech", *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 19, no. 4, pp. 947–960, Sep. 2010. https://doi.org/10.1109/TASL.2010.2072499

[12] S. Skodda, "Aspects of speech rate and regularity in parkinson's disease", *J Neurol Sci*, vol. 310, no. 1-2, pp. 231–236, Aug. 2011. https://doi.org/10.1016/j.jns.2011.07.020

[13] T. Villa-Cañas, J. Orozco-Arroyave, J. Vargas-Bonilla & J. Arias-Londoño, "Modulation spectra for automatic detection of parkinson's disease", In Proc. *2014 XIX STSIVA*, IEEE, AM, CO, Sept. 17-19, 2014, pp. 1–5. https://doi.org/10.1109/STSIVA.2014.7010173

[14] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva & E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease", in Proc. *LREC´14*, ELRA, RKV, ISL, May. 26-31, 2014, pp. 342–347. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/7_Paper.pdf

[15] J. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz & E. Nöth, "Automatic detection of parkinson's disease in running speech spoken in three different languages", *JASA*, vol. 139, no. 1, pp. 481–500, 2016. https://doi.org/10.1121/1.4939739

[16] H. Ackermann, I. Hertrich & T. Hehr, "Oral diadochokinesis in neurological dysarthrias", *Folia Phoniatr Logop*, vol. 47, no. 1, pp. 15–23, Feb. 1995. https://doi.org/10.1159/000266338

[17] C.-C. Yang, Y.-M. Chung, L.-Y. Chi, H.-H. Chen & Y.-T. Wang, "Analysis of verbal diadochokinesis in normal speech using the diadochokinetic rate analysis program", *JDS*, vol. 6, no. 4, pp. 221–226, Dec. 2011. https://doi.org/10.1016/j.jds.2011.09.007

[18] M. N. Wong, B. E. Murdoch & B.-M. Whelan, "Lingual kinematics during rapid syllable repetition in parkinson's disease", *Int J Lang Commun Disord*, vol. 47, no. 5, pp. 578–588, Jul. 2012. https://doi.org/10.1111/j.1460-6984.2012.00167.x

[19] M. Lotze, G. Seggewies, M. Erb, W. Grodd & N. Birbaumer, "The representation of articulation in the primary sensorimotor cortex", *Neuroreport*, vol. 11, no. 13, pp. 2985–2989, Sep. 2000. https://doi.org/10.1097/00001756-200009110-00032

[20] D. O'Shaughnessy, *Speech Communications: Human and Machine*, CAM, USA: Universities Press, 2009.

[21] D. Montaña, Y. Campos-Roca & C. J. Pérez, "A diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of parkinson's disease", *Comput Meth Prog Bio*, vol. 154, pp. 89–97, Feb. 2018. https://doi.org/10.1016/j.cmpb.2017.11.010

[22] M. Novotný, J. Rusz, R. Čmejla & E. Růžička, "Automatic evaluation of articulatory disorders in parkinson's disease", *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014. https://doi.org/10.1109/TASLP.2014.2329734

[23] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. V. Tyagi & C. Wellekens, "Automatic speech recognition and speech variability: A review", *Speech Commun*, vol. 49, no. 10-11, pp. 763–786, Oct. 2007. https://doi.org/10.1016/j.specom.2007.02.006

[24] J. Markič, "Real academia española y asociación de academias de la lengua española. Nueva gramática de la lengua española. Fonética y fonología", *Linguistica*, vol. 52, no. 1, pp. 403–406, Dec. 2012. https://doi.org/10.4312/linguistica.52.1.403-406

[25] A. M. Borzone, *Manual de fonética acústica*, TX, USA: Hachette, 1980.

[26] J. Rusz, R. Cmejla, H. Ruzickova & E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease", *JASA*, vol. 129, no. 1, pp. 350–367, Feb. 2011. https://doi.org/10.1121/1.3514381

[27] L. Rabiner & R. Schafer, *Digital processing of speech signals.* ENGL, N.J., USA: Prentice-Hall, 1978.

[28] J. Proakis & D. Manolakis, *Digital signal processing: principles, algorithms, and applications.* USR, N.J., USA: Prentice Hall, 1996.

[29] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges", *Pattern Recognit. Image Anal.*, vol. 41, no. 10, pp. 2965–2979, Oct. 2008. https://doi.org/10.1016/j.patcog.2008.05.008

[30] Q. Jin & T. F. Zheng, "Overview of front-end features for robust speaker recognition", in Proc. *ASC2011*, APSIPA, Xi'an, CN, Oct. 18-20, 2011. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.704.6205&rep=rep1&type=pdf

[31] P. Maragos, J. F. Kaiser & T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis", *IEEE Trans. Signal Process*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993. https://doi.org/10.1109/78.277799

[32] M. Grimaldi & F. Cummins, "Speaker identification using instantaneous frequencies", *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 16, no. 6, pp. 1097–1111, Aug. 2008. https://doi.org/10.1109/TASL.2008.2001109

[33] S. O. Sadjadi & J. H. L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions", in Proc. *INTERSPEECH2010*, ISCA, Makuhari, JP, Sep. 26-30, 2010, pp. 2138–2141. Available at: https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_2138.pdf

[34] P. Clark & L. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals", *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4323–4332, Nov. 2009. https://doi.org/10.1109/TSP.2009.2025107

[35] M. Sarria-Paja & T. Falk, "Whispered speech detection in noise using auditory-inspired modulation spectrum features", *IEEE Signal Process Lett*, vol. 20, no. 8, pp. 783–786, Aug. 2013. https://doi.org/10.1109/LSP.2013.2266860

[36] T. Falkand & W.-Y. Chan, "Modulation spectral features for robust farfield speaker identification", *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 18, no. 1, pp. 90–100, Jan. 2010. https://doi.org/10.1109/TASL.2009.2023679

[37] S. Ganapathy, S. Thomas & H. Hermansky, "Static and dynamic modulation spectrum for speech recognition", in Proc. *INTERSPEECH2009*, ISCA, Brig, UK, Sep. 6-10, 2009, pp. 2823–2826. Available at: https://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_2823.pdf

[38] T. Kinnunen & H. Li, "An overview of text-independent speaker recognition: From features to supervectors", *Speech Commun*, vol. 52, no. 1, pp. 12–40, Jan. 2010. https://doi.org/10.1016/j.specom.2009.08.009

[39] M. Sarria-Paja & T. H. Falk, "Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification", *Comp Speech Lang*, vol. 45, pp. 437–456, Sep. 2017. https://doi.org/10.1016/j.csl.2017.04.004

[40] C. Bishop, *Pattern Recognition and Machine Learning.* NY, USA: Springer-Verlag, 2006.

[41] M. Senoussaoui, M. Saria-Paja, P. Cardinal, T. H. Falk & F. Michaud, "1. State-of-the-art speaker recognition methods applied to speakers with dysarthria" in *Voice Technologies for Speech Reconstruction and Enhancement*, H. A. Patil & A. Neustein, Ed. BE, GE: De Gruyter, 2020. pp. 7–34. https://doi.org/10.1515/9781501501265-002

[42] O. Tuzel, F. Porikli & P. Meer, "Region covariance: A fast descriptor for detection and classification", in Proc. *ECCV2006*, GRZ, AT, May. 7-13, 2006, pp. 589–600. https://doi.org/10.1007/11744047_45

[43] O. Tuzel, F. Porikli & P. Meer, "Human detection via classification on riemannian manifolds", In Proc. *CVPR'07*, IEEE, Mpls, MN, USA, Jun. 17-22, 2007, pp. 1–8. https://doi.org/10.1109/CVPR.2007.383197

[44] C. Ye, J. Liu, C. Chen, M. Song & J. Bu, "Speech Emotion Classification on a Riemannian Manifold", in Proc. *9th PCM2008*, LNCS, TNN, TW, Dec. 9-13, 2008, , vol. 5353, pp. 61–69. https://doi.org/10.1007/978-3-540-89796-5

[45] R. Duda, P. Hart & D. Stork, *Pattern classification*, NY, USA: John Wiley & Sons, 2012.

[46] L. Rabiner & B. Juang, "An introduction to hidden markov models", *IEEE ASSP Mag*, vol. 3, no. 1, pp. 4–16, Jan. 1986. https://doi.org/10.1109/MASSP.1986.1165342

[47] V. N. Vapnik, "An overview of statistical learning theory", *IEEE Trans Neural Netw Learn Syst*, vol. 10, no. 5, pp. 988–999, Sep. 1999. https://doi.org/10.1109/72.788640

[48] B. Scholkopfand & A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, CAM, USA: MIT Press, 2001.

[49] R. Lyon, A. Katsiamis & E. Drakakis, "History and future of auditory filter models", in Proc. *ISCAS*, IEEE, PAR, FRA, May. 30-Jun. 2, 2010, pp. 3809–3812. https://doi.org/10.1109/ISCAS.2010.5537724

[50] T. Cho & P. Ladefoged, "Variation and universals in vot: evidence from 18 languages", *J. Phon*, vol. 27, no. 2, pp. 207–229, Apr. 1999. https://doi.org/10.1006/jpho.1999.0094

[51] L. Liskerand & A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements", *Word*, vol. 20, no. 3, pp. 384–422, 1964. https://doi.org/10.1080/00437956.1964.11659830

Jhon F. Moofarry-Villaquiran, received the Bs. Eng in Electronic Engineering in 2019. He worked projects in signal processing, robotics and since 2018 with emphasis on signal processing for the Universidad Santiago de Cali in the publication of scientific journals. He is currently working on a project about deep learning. https://orcid.org/0000-0002-0366-5396

Patricia Argüello-Vélez, received the degree in speech therapist from the University of Santiago de Cali, Valle, Colombia, in 2007, and master's degree in Linguistics from the Pereira technology university, Pereira, Colombia, in 2013. Currently, she is a PhD candidate in linguistics at the University of Antioquia. She is professor at the Faculty of Health at the University of Santiago de Cali, member of the research group in speech therapy and psychology of the university of Santiago de Cali and member of the Sociolinguistic Studies Group of University of Antioquia. https://orcid.org/0000-0002-5733-3506

Milton Sarria-Paja, received the Bs. Eng in Electronic Engineer and Master in Industrial Automation from the National University of Colombia, in 2006 and 2009, respectively. He received his PhD degree in 2017 from the Institut National de la Recherche Scientifique (INRS-EMT) University of Quebec, Montreal, QC, Canada. He currently works as a full-time professor for the Faculty of Engineering of the Universidad Santiago de Cali-Colombia. His areas of interest correspond to Artificial Intelligence using Machine Learning and Signal Processing, and its relationship with mathematics and statistics. https://orcid.org/0000-0003-4288-1742