TecnoLógicas
ISSN-p 0123-7799
ISSN-e 2256-5337
Vol. 23, No. 47, pp. 93-108
Enero-abril de 2020

Artículo de Revisión/Research Article

Cepstral Analysis and Hilbert-Huang Transform for Automatic Detection of Parkinson's Disease

Análisis cepstral y la transformada de Hilbert-Huang para la detección automática de la enfermedad de Parkinson

Felipe O. López-Pabón[1],
Tomas Arias-Vergara[2] y
Juan R. Orozco-Arroyave[3]

Recibido: 26 de junio de 2019
Aceptado: 04 de octubre de 2019

Cómo citar / How to cite
F. O. López-Pabón, T. Arias-Vergara, J. R. Orozco-Arroyave, "Cepstral Analysis and Hilbert-Huang Transform for Automatic Detection of Parkinson's Disease", *TecnoLógicas*, vol. 23, no. 47, pp. 93-108, 2020. https://doi.org/10.22430/22565337.1401

[1] Ingeniero Electrónico, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia, forlando.lopez@udea.edu.co

[2] MSc. en Ingeniería, Facultad de Ingeniería, Universidad de Antioquia, Laboratorio de reconocimiento de patrones (LME), Medellín-Colombia, Universidad de Erlangen-Núremberg, Erlangen-Alemania, Universidad de Múnich, Múnich-Alemania, tomas.arias@udea.edu.co

[3] PhD. en Ciencias de la Computación, Grupo de investigación en Telecomunicaciones aplicadas (GITA), Facultad de Ingeniería, Universidad de Antioquia, Laboratorio de reconocimiento de patrones (LME), Medellín-Colombia, Universidad de Erlangen-Núremberg, Erlangen-Alemania, rafael.orozco@udea.edu.co

Abstract

Most patients with Parkinson's Disease (PD) develop speech deficits, including reduced sonority, altered articulation, and abnormal prosody. This article presents a methodology to automatically classify patients with PD and Healthy Control (HC) subjects. In this study, the Hilbert-Huang Transform (HHT) and Mel-Frequency Cepstral Coefficients (MFCCs) were considered to model modulated phonations (changing the tone from low to high and vice versa) of the vowels /a/, /i/, and /u/. The HHT was used to extract the first two formants from audio signals with the aim of modeling the stability of the tongue while the speakers were producing modulated vowels. Kruskal-Wallis statistical tests were used to eliminate redundant and non-relevant features in order to improve classification accuracy. PD patients and HC subjects were automatically classified using a Radial Basis Support Vector Machine (RBF-SVM). The results show that the proposed approach allows an automatic discrimination between PD and HC subjects with accuracies of up to 75 % for women and 73 % for men.

Keywords

Speech articulation, Classification, Hilbert-Huang, Parkinson's Disease.

Resumen

La mayoría de las personas con la enfermedad de Parkinson (EP) desarrollan varios déficits del habla, incluyendo sonoridad reducida, alteración de la articulación y prosodia anormal. Este artículo presenta una metodología que permite la clasificación automática de pacientes con EP y sujetos de control sanos (CS). Se considera que la transformada de Hilbert-Huang (THH) y los Coeficientes Cepstrales en las frecuencias de Mel modelan las fonaciones moduladas (cambiando el tono de bajo a alto y de alto a bajo) de las vocales /a/, /i/, y /u/. La THH se utiliza para extraer los dos primeros formantes de las señales de audio, con el objetivo de modelar la estabilidad de la lengua mientras los hablantes producen vocales moduladas. Pruebas estadísticas de Kruskal-Wallis se utilizan para eliminar características redundantes y no relevantes, con el fin de mejorar la precisión de la clasificación. La clasificación automática de sujetos con EP vs. CS se realiza mediante una máquina de soporte vectorial de base radial. De acuerdo con los resultados, el enfoque propuesto permite la discriminación automática de sujetos con EP vs. CS con precisiones de hasta el 75 % para los hombres y 73 % para las mujeres.

Palabras clave

Articulación del habla, clasificación, Hilbert-Huang, enfermedad de Parkinson.

1. INTRODUCTION

Parkinson's Disease (PD) is a progressive neurodegenerative condition that affects approximately 2 % of the population over 65 years old [1].

Individuals with PD usually present motor deficits, such as tremor, rigidity, akinesia, bradykinesia, and postural instability. Additionally, more than 90 % of said patients develop several speech deficits such as reduced sonority, monotonicity, imprecise articulation, and abnormal prosody [2]-[4]. Most speech studies with PD patients evaluate sustained vowel phonations because it is one of the easiest tasks compared with monologues or reading long texts, and provides valuable information about phonation and articulation dimensions of speech production [5].

Several studies have found articulation deficits in PD patients using sustained vowel phonations and continuous speech recordings. For instance, in [6], the authors evaluated the characteristics of the articulation of vowels pronounced by 35 native Czech speakers (20 with PD and 15 HC) while performing continuous speech tasks. According to their results, impaired vowel articulation may be considered a possible early biomarker of PD. They stressed the fact that vowel articulation problems may be evaluated in continuous speech signals to obtain accurate models of speech impairments in PD patients. They also reported classification accuracies of up to 80 % in the automatic discrimination between PD and HC subjects.

Other studies have considered articulatory acoustic features extracted from vowels. In [7], the authors considered speech recordings of 68 PD patients and 32 HC (native German speakers). The first two formant frequencies (F1 and F2) were extracted from the vowels /a/, /i/, and /u/, which were segmented from continuous speech recordings. An articulatory acoustic analysis was performed using the Triangular Vowel Space Area (tVSA) and the Vowel Articulation Index (VAI). Said authors reported VAI values significantly reduced in male and female PD patients compared to the HC group. tVSA was only reduced in male PD patients. Therefore, they concluded that VAI seemed to be more efficient than tVSA to identify articulation deficits of PD patients. These measurements have also been considered in recent studies that asses the automatic evaluation of articulation deficits of PD patients observed in sustained vowels and continuous speech signals [8].

Although the conclusions and observations in the literature are well motivated and supported on strong arguments, it is important to note that most studies in the field have considered classical sustained vowels and continuous speech signals, but not other specific tasks like the production of modulated vowels (i.e., changing the tone from low to high and from high to low). We think that this kind of tasks could be more accurate and robust than classical ones to observe specific speech deficits in PD patients because they are easy to perform and allow the measurement of frequency variations in the voice. For instance, modulated vowels were considered in [5], where recordings of 50 patients with PD and 50 HC subjects (all of them Colombian Spanish native speakers) were discriminated. The feature set included the classical vocal formants and the fundamental frequency extracted using the Hilbert-Huang Transform (HHT).

The automatic discrimination of PD vs. HC speakers was performed implementing a decision tree classifier. The classification experiments were conducted considering male and female speakers separately. The accuracies obtained with the classical sustained phonations were around 82 % and 90 %, respectively; when the modulated vowels were included, the results improved to 84 %.

The analysis provided by the HHT has been used in multiple papers. For example, HHT was used for a geophysical study about the propagation of seismic waves [9].

They concluded that certain Intrinsic Mode Functions (IMFs), those with a higher frequency, could be identified as generated near the hypocenter, while the high frequency content was related to a large tension drop associated with the onset of seismic events. In [10], they used the HHT to study financial time series and as a tool for the statistical analysis of nonlinear and non-stationary data. They obtained space time-frequency decompositions and temporal decompositions of the data.

In [5], the HHT was implemented to compute the instantaneous energy and its range, as well as the instantaneous frequency and the difference between the maximum and minimum values of the IMFs amplitude. In turn, in this study, we used the Hilbert-Huang transform to extract the first and second IMF, which encode information about the temporal variation of the vocal formants (F1 and F2). We extracted features from such IMFs as described in Section 2.3.

In this paper, we introduce the use of HHT for a robust modeling of the frequency bands where the first two vocal formants are located because these two are the main determinants of which vowel is heard, and, in general, they are responsible for the differences in quality between different periodic sounds.

The modeling was based on the extraction of the IMFs that result from the Empirical Mode Decomposition (EMD).

Such frequency bands around the formants were modeled along with their energy content, their first and second derivatives, the instantaneous frequency, their Teager Energy operator (TEO) value, and their entropy. Besides these features, the classical Mel-Frequency Cepstral Coefficients (MFCCs) were considered.

This study aims to contribute with a novel, original, and robust alternative to model articulatory deficits exhibited by PD patients. The experiments employed recordings of the vowels /a/, /i/, and /u/ pronounced in a modulated tone. The same set of recordings of [5] was considered in this study to evaluate the capability of the proposed approach to discriminate between PD and HC speakers and evaluate the severity of their dysarthria.

This paper is organized as follows: Section 2 describes the database, the methodology, the way the features were extracted, and the implemented algorithms. Section 3 details the experiments and discusses their results.

Finally, Section 4 presents the conclusions and future work.

2. MATERIALS AND METHODS

2.1 Participants

Recordings of the corpus PC-GITA were considered here. A total of 100 participants were included, 50 with PD and 50 HC [11].

All the participants were Colombian Spanish native speakers. Their clinical and demographic information is provided in Table 1. The recordings were captured with a sampling frequency of 44.1 KHz and a 16-bit resolution. The participants were asked to pronounce sustained modulated vowels (/a/, /i/, and /u/) during one single breath. All the patients were evaluated by an expert neurologist according to the Movement Disorder Society-Unified Parkinson's Disease. Rating Scale (MDS-UPDRS) [12].

Table 1. Clinical and demographic information of the participants. Source: Created by the authors.

	PD Patients		**HC Speakers**	
	Male	Female	Male	Female
Number of subjects	25	25	25	25
Age [years] (μ ± σ)	61.6 ± 11.6	60.6 ±7.3	62.6 ± 9.3	61.4 ± 6.9
Age range [years]	33–81	49–75	42–86	49–76
Time after diagnosis [years] (μ ± σ)	8.9 ± 5.9	12.6 ± 11.5	-	-
MDS-UPDRS-III (μ ± σ)	37.7 ± 21.9	37.5 ± 13.9	-	-
MDS-UPDRS-III range	6–92	19–71	-	-
m-FDA(μ ± σ)	28.9 ± 8.4	26.9 ± 8.3	8.7 ± 6.6	6.6 ± 7.0
m-FDA range	13–41	13–47	0–25	0–23

PD: Parkinson's disease, **HC:** Healthy controls, μ: average, σ: standard deviation.

Additionally, with the aim of obtaining a label for the dysarthria level of the participants, all the recordings were evaluated by three phoniatricians according to the modified version of the Frenchay Dysarthria Assessment tool (m-FDA). Further information about the procedure and the scale can be found in [13].

2.2 Methodology

The methodology followed in this study is shown in Fig. 1. The procedure begins with the pre-processing of the audio file, which consists of eliminating the DC level of the signal and normalizing the amplitude. The second step is the feature extraction, followed by the classification, which, in this case, is performed using a Support Vector Machine (SVM).

The performance of the system is evaluated using different statistics including accuracy (Acc), sensitivity (Sens), specificity (Spec), receiver Operating Characteristic Curve (ROC), and the Area Under the ROC Curve (AUC). Further details about each step are provided below.

2.3 Feature Extraction

Hilbert-Huang Transform: For an arbitrary time series, $X(t)$, it is always possible to calculate its Hilbert transform, $Y(t)$, as (1):

$$Y(t) = \frac{1}{\pi} P \int \frac{X(t')}{t - t'} dt' \tag{1}$$

where, P indicates the principal value of the Cauchy integral, and t and t' are two different time instants. With this definition, $X(t)$ and $Y(t)$ form a complex conjugate pair. Hence, it is possible to obtain the following analytic signal, $Z(t)$ (2):

$$Z(t) = X(t) + iY(t) = \mathrm{a(t)} e^{i\theta(t)} \tag{2}$$

such that

Note that (1) defines the Hilbert Transform as the convolution between $X(t)$ and $1/t$. Therefore, it emphasizes the local properties of $X(t)$, although the transformation is global. In (2), the expression of polar coordinates further clarifies the local nature of this representation, which shows the local adjustment of a trigonometric function that varies in amplitude and phase to $X(t)$ (3).

$$a(t) = [X^2(t) + Y^2(t)]^{\frac{1}{2}}$$
$$\theta(t) = \arctan\left[\frac{Y(t)}{X(t)}\right] \tag{3}$$

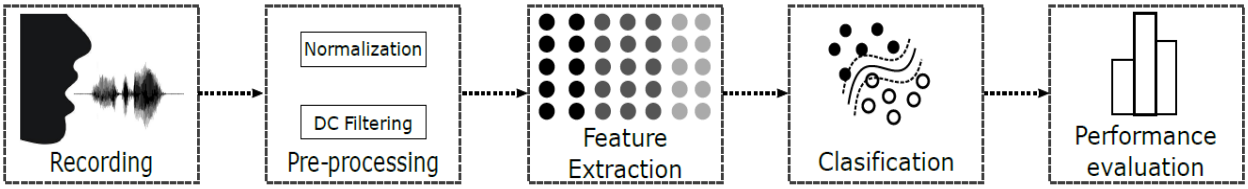

Fig. 1. Block diagram of the methodology implemented in this study. Source: Created by the authors.

One of the advantages of the Hilbert transform became popular after Huang et.al. in [14], where the EMD method and the use of HT were introduced. EMD is necessary to pre-process the data before applying the HT. EMD also reduces the data to a collection of IMFs, which are defined as functions that satisfy the following conditions: (i) the number of extremes and zero crossings are the same or differ by a maximum of one over all the data, and (ii) the average value of the envelope defined by the local maxima and the envelope defined by the local minima is zero at any point. EMD decomposes an arbitrary and time-varying signal into IMFs that are modulated in amplitude and frequency [5]. Those IMFs represent the frequencies that are present in the signal, and the sum of those functions reconstructs the original signal.

In speech, the resonant frequencies of the vocal tract are called formants. The most informative formants are the first two, F1 and F2. According to [15], the information of F1 and F2 per vowel is located around the following frequency bands: /a/, between 600 Hz and 1700 Hz; /i/, between 200 Hz and 2600 Hz; and /u/, between 200 Hz and 1100 Hz. Given the capability of IMFs to separate relevant information in different frequency bands, in this study, we take the original band-pass filtered signals and extract their corresponding IMFs to automatically model information of F1 and F2. Note that only one band-pass filter is used per signal, which reduces the risk of introducing alias frequencies or the problem of leakage.

Fig. 2. shows the wave forms of the vowel /a/ produced with a modulated tone by a 63-year-old healthy control female (left) and by a 55-year-old female PD patient (right). The patient was diagnosed 12 years ago and the label of her MDS-UPDRS score is 43. The spectrograms of the original signals show the modulation of their frequency content. Fig. 2B, Fig. 2C, Fig. 2E, and Fig. 2F show the IMF signals and their corresponding spectrograms after applying the filtering around the frequencies of the vocal formants. The first IMF allows the modeling of the first formant (Fig. 2B and Fig. 2E), while the second formant is modeled by the second IMF (Fig. 2C and Fig. 2F). Note that the first two IMFs obtained from the patient (Fig. 2E and Fig. 2F) are jumpier, which indicates the possible presence of vocal tremor, which is one of the most common behaviors in PD patients and is typically linked to the difficulty of patients to produce stable vowel phonations.

Considering that the HHT provides a high resolution in the frequency domain (especially in low-frequency bands), which is even better than that of the Fourier transform and the Wavelet transform [14], this method is useful when more detailed information of the modulated vowels is required.

Several characteristics of the 2 extracted IMFs are measured: energy content, first and second derivatives, instantaneous frequency computed in windows of 20 ms with a step size of 10 ms, Teager energy operator (TEO), and entropy (also considered for the audio signal).

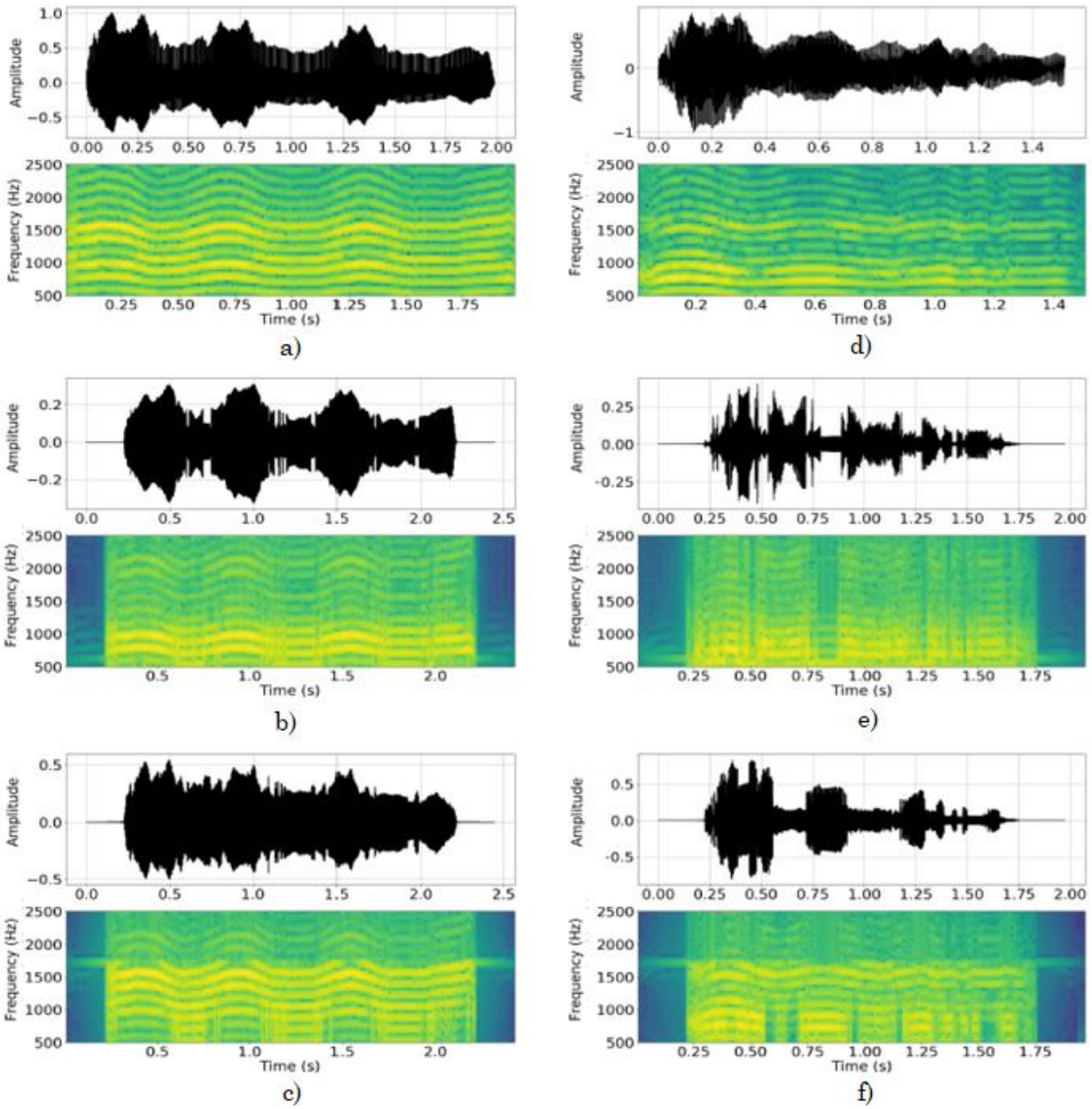

Fig. 2. Waveforms of modulated vowel /a/ and their corresponding spectrogram. (a) Original audio signal of a 63-year-old female HC, (b) first IMF of the HC signal, (c) second IMF of the HC signal, (d) original audio signal of a 55-year-old female PD patient with 43 points in the MDS-UPDRS-III scale and 12 years after diagnosis, (e) first IMF of the PD signal, (f) second IMF of the PD signal. Source: Created by the authors.

Mean value, standard deviation, skewness, and kurtosis are calculated per measurement to create a 32-dimensional feature vector per speaker.

Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are very common in speech processing due to their robustness and suitability for several applications such as speech recognition, speaker verification, and speaker identification. They are based on the critical band model of sound perception, which is emulated by a filter bank with band spacing and bandwidth similar to the critical bands given by the Mel scale. The conversion from Hertz to Mels is shown in (4).

As MFCCs are based on human perception, they have been successfully used to model articulation in PD [16]. The steps to calculate MFCCs are the following: 1) a short-term segment of the signal is extracted, 2) a window (e.g., Hamming) is applied upon the segment, 3) Fourier analysis is performed by the Discrete Fourier Transform, 4) a triangular filter bank is applied to the DFT to estimate the Mel Energy spectrum, 5) the natural logarithm of Mel's energy spectrum is calculated, and 6) the discrete cosine transform (DCT) is calculated to obtain the MFCC [17].

In this study, we are considering only the first 13 MFCCs as well as their first and second derivatives. From coefficient 14 onwards, the information about the phenomenon under analysis (human voice) is irrelevant. The coefficients are extracted from 30-ms-long frames (approximately statistically stationary signals), with an overlap of 10 ms. Similar to the model of the IMFs, mean value, standard deviation, skewness, and kurtosis are calculated to create a 156-dimensional feature vector per speaker.

2.4 Feature Selection

Given that the number of extracted features is relatively high, we want to evaluate whether a lower dimensional representation is more suitable to perform the automatic classification of PD patients and HC subjects. The representation space can be reduced via dimensionality reduction strategies such as those based on Principal Component Analysis or Linear Discriminant Analysis, or via feature selection. In this work, we decided to evaluate the suitability of the second one.

Particularly, we applied Kruskal-Wallis tests to evaluate the null hypothesis of a group with n independent samples that come from the same population or from identical populations with the same median. To determine whether any of the differences between the medians is statistically significant, the p value is compared with the level of significance to evaluate the null hypothesis. The null hypothesis indicates that the population averages are all the same. In general, a level of significance (namely α) of 0.05 works properly. α = 0.05 indicates a 5 % risk of concluding that there is a difference when there is no real difference between the two medians. When $p \leq \alpha$, the null hypothesis is rejected, and it is concluded that not all the population medians are equal. In turn, when $p \geq \alpha$, there is not enough evidence to reject the null hypothesis that the population medians are all equal. A level of significance α = 0.05 was established in our experiments to exclude redundant features that do not provide relevant information for the discrimination between classes.

$$Mel = 2595 \log_{10}\left(1 + \frac{f}{7000}\right) = 1127 \ln\left(1 + \frac{f}{700}\right) \quad (4)$$

2.5 Automatic classification

To differentiate between patients and healthy speakers, an SVM with Gaussian kernel was considered. Parameters C and ɣ were optimized through a grid search up to powers of ten with C ∈ {0.001, 0.01, …, 1000, 10000} and ɣ ∈ {0.0001, 0.001, …, 100, 1000}. A speaker independent 10-fold Cross-Validation (CV) strategy was implemented in the training process, i.e., the data were divided into 10 groups (randomly chosen); 9 of them were used for training and the remaining one for the test. The same approach was adopted in [5], where the same corpus was used in the experiments.

3. EXPERIMENTS AND RESULTS

Two cases are considered in this study: (1) classification of PD vs. HC without discriminating by gender of the participants, and (2) classification of PD vs. HC for men and women separately. The experiments were carried out taking into account three groups of features: the features obtained from HHT (G1), the features with MFCCs (G2), and the combination of G1 and G2 (G3). The analysis of the three vowels (/a/, /i/, and /u/) was carried out separately and the 3 vowels together. This is because the vocal space area (VSA) formed by these vowels provides valuable information about the impact of PD in voice generation, and it quantifies the possible reduction in the articulatory capability of the speaker [8].

3.1 Bi-class classification

Table 2 shows the results of the classification between PD vs. HC subjects without any prior sex-based grouping. In G1 and G2, the best result was obtained from vowel /i/; meanwhile, in G3, the best result was obtained considering vowel /a/.

Finally, the highest accuracy, 69 %, was obtained with optimal parameters C = 10 and ɣ = 1.10^{-2}. Note that the feature selection process improved the accuracy of the model. Despite the feature selection, the results did not exceed a 70 % accuracy.

We think that it is because a prior grouping based on the gender of the speakers is required due to the fact that women produce more aspiration noise than men (which affects spectral regions that correspond to first formants [18]) and there are differences between the low-pitched voices of men and the high-pitched voices of women. This is reported below.

The results of the classification, considering a prior gender-based grouping of speakers, are reported in Table 3 and Tabla 4.

Note that the results are slightly better than those obtained in the previous experiment. Relatively high accuracies were achieved, especially when the feature selection was carried out with the Kruskal-Wallis test. Remarkably, the highest accuracies, in general, were obtained with vowel /i/, which confirms that this vowel contributes to the computation of VSA with suitable information to quantify the reduction of the articulatory capability of PD speakers.

When the original feature vector is taken into account, for men and women, in general, the best result was obtained with the vowel /i/. When the feature selection was performed with the Kruskal-Wallis test, the best results were obtained with vowels /i/ and /a/ and the combination of /a/, /i/, and /u/. Better results were obtained for female than male participants (accuracies of 75 % and 73 %, respectively). This can be explained because, compared to men with PD, women with PD suffer more notoriously from a marked subharmonic energy and segments with voice breaks [19].

Fig. 3 and Fig. 4 show the scores of the classifier, which refer to the distance of each sample to the separating hyperplane.

Note that, in general, the scores overlap and the scores of patients with PD are more equally dispersed than those of the HC group. Therefore, there is a tendency to confuse patients and healthy controls.

Among other factors, this is due to the fact that the age of both populations is very similar and parameter c (which controls the distance to the hyperplane between classes), in general, is a very low value.

However, it should be clarified that patients with PD and HC speakers can be (slightly) better classified using the feature selection with Kruskal-Wallis test.

Table 2. Results of PD vs. HC speakers without classification by sex
Source: Created by the authors.

		Original vector					**With feature selection**				
		C	**ɣ**	**Acc (%)**	**Sen (%)**	**Spe (%)**	**C**	**ɣ**	**Acc (%)**	**Sen (%)**	**Spe (%)**
G1	/a/	1	0.01	**68 ± 3**	62 ± 5	74 ± 3	1	0.01	**67 ± 2**	59 ± 3	75 ± 3
	/i/	1	1	63 ± 3	62 ± 7	65 ± 7	1	0.01	65 ± 4	73 ± 4	58 ± 8
	/u/	1	1	62 ± 2	75 ± 4	49 ± 5	1	0.01	63 ± 3	73 ± 3	54 ± 4
	Fusion	0.5	1	63 ± 2	59 ± 4	66 ± 3	1	0.01	65 ± 3	65 ± 6	65 ± 5
G2	/a/	1000	1	65 ± 1	60 ± 4	70 ± 4	1	0.01	64 ± 3	59 ± 3	70 ± 6
	/i/	5	1	**67 ± 4**	68 ± 4	66 ± 7	5	0.01	**68 ± 4**	66 ± 6	69 ±4
	/u/	5	0.01	61 ± 3	64 ± 3	57 ± 9	1	0.01	63 ± 3	58 ± 7	67 ± 7
	Fusion	5	1	63 ± 4	61 ± 9	66 ± 6	10	0.01	65 ± 3	62 ± 6	68 ± 4
G3	/a/	5	1	63 ± 2	55 ± 5	72 ± 4	1	0.01	68 ± 2	65 ± 5	71 ± 4
	/i/	1	1	**69 ± 3**	68 ± 4	69 ± 3	10	0.01	**69 ± 3**	69 ± 5	68 ± 6
	/u/	10	1	62 ± 4	64 ± 6	60 ± 5	1	0.01	63 ± 3	65 ± 6	61 ± 6
	Fusion	0.001	0.0001	65 ± 2	64 ± 4	65 ± 4	5	0.001	66 ± 2	69 ± 3	63 ± 3

G1: Features obtained from HHT. G2: Features with MFCCs. G3: Fusion of G1 and G2. **Acc:** Accuracy. **Sen:** Sensitivity. **Spe:** Specificity. **Fusion:** Merging of the three vowels.

Table 3. Results of male PD patients vs. male HC speakers
Source: Created by the authors.

		Original vector					**With feature selection**				
		C	**ɣ**	**Acc (%)**	**Sen (%)**	**Spe (%)**	**C**	**ɣ**	**Acc (%)**	**Sen (%)**	**Spe (%)**
G1	/a/	0.001	0.0001	65 ± 4	46 ± 8	84 ± 4	0.001	0.01	69 ± 4	56 ± 7	82 ± 5
	/i/	5	0.001	**68 ± 4**	68 ± 6	67 ± 9	0.001	0.0001	**73 ± 4**	74 ± 6	72 ± 8
	/u/	0.001	0.001	66 ± 5	73 ± 8	59 ± 14	5	0.01	66 ± 4	73 ± 9	58 ± 11
	Fusion	50	0.0001	63 ± 6	64 ± 10	62 ± 8	5	0.01	71 ± 4	64 ± 4	78 ± 7
G2	/a/	5	0.001	64 ± 6	63 ± 7	65 ± 11	0.001	0.1	65 ± 3	63 ± 10	68 ± 10
	/i/	0.001	0.01	**66 ± 3**	73 ± 8	58 ± 10	0.001	0.01	63 ± 4	64 ± 9	63 ± 7
	/u/	0.001	0.01	62 ± 4	72 ± 7	52 ± 9	0.001	0.01	65 ± 3	62 ± 7	68 ± 8
	Fusion	0.001	0.0001	63 ± 4	63 ± 8	63 ± 5	5	0.01	**68 ± 4**	71 ± 5	66 ± 5
G3	/a/	0.001	1	62 ± 4	40 ± 9	84 ± 8	1	0.01	67 ± 4	63 ± 10	71 ± 7
	/i/	0.001	0.0001	**69 ± 2**	74 ± 4	65 ± 6	1	0.01	**73 ± 4**	76 ± 6	71 ± 5
	/u/	0.001	0.01	65 ± 3	74 ± 7	56 ± 10	0.001	0.01	67 ± 4	67 ± 6	68 ± 6
	Fusion	0.001	0.1	61 ± 6	47 ± 10	75 ± 8	5	0.001	68 ± 4	69 ± 5	67 ± 6

G1: Features obtained from HHT. G2: Features with MFCCs. G3: Fusion of G1 and G2.
Acc: Accuracy. **Sen:** Sensitivity. **Spe:** Specificity. **Fusion:** Merging of the three vowels.

Table 4. Results of female PD patients vs. female HC speakers
Source: Created by the authors.

		Original vector					With feature selection				
		C	**ɣ**	**Acc (%)**	**Sen (%)**	**Spe (%)**	**C**	**ɣ**	**Acc (%)**	**Sen (%)**	**Spe (%)**
G1	/a/	0.001	0.01	**70 ± 2**	59 ± 5	81 ± 6	0.001	0.0001	71 ± 5	65 ± 10	75 ± 8
	/i/	0.001	0.001	66 ± 4	64 ± 8	68 ± 7	0.001	0.01	69 ± 4	67 ± 7	70 ± 5
	/u/	0.001	0.001	67 ± 4	71 ± 6	64 ± 9	0.001	0.1	69 ± 4	73 ± 7	65 ± 3
	Fusion	0.001	0.001	68 v 4	63 ± 6	74 ± 6	0.001	0.0001	**73 ± 2**	73 ± 6	72 ± 6
G2	/a/	5	0.001	63 ± 3	72 ± 8	55 ± 5	0.001	0.1	64 ± 5	64 ± 11	64 ± 10
	/i/	5	0.001	**71 ± 3**	70 ± 7	73 ± 5	5	0.001	**75 ± 4**	71 ± 7	78 ± 7
	/u/	50	0.001	62 ± 5	57 ± 12	67 ± 9	0.001	0.01	67 ± 7	64 ± 9	70 ± 8
	Fusion	10	0.0001	67 ± 4	69 ± 8	65 ± 9	0.001	0.01	69 ± 4	62 ± 7	76 ± 3
G3	/a/	1	0.01	64 ± 4	62 ± 8	67 ± 5	0.001	0.01	**70 ± 5**	67 ± 8	73 ± 6
	/i/	10	0.0001	**72 ± 4**	70 ± 6	73 ± 5	0.001	0.01	69 ± 3	67 ± 6	71 ± 6
	/u/	0.001	1	64 ± 4	60 ± 9	68 ± 8	0.001	0.01	68 ± 3	64 ± 7	72 ± 5
	Fusion	0.001	0.0001	66 ± 5	66 ± 8	66 ± 9	0.001	0.0001	67 ± 3	68 ± 5	67 ± 7

G1: Features obtained from HHT. G2: Features with MFCCs. G3: Fusion of G1 and G2.
Acc: Accuracy. **Sen:** Sensitivity. **Spe:** Specificity. **Fusion:** Merging of the three vowels.

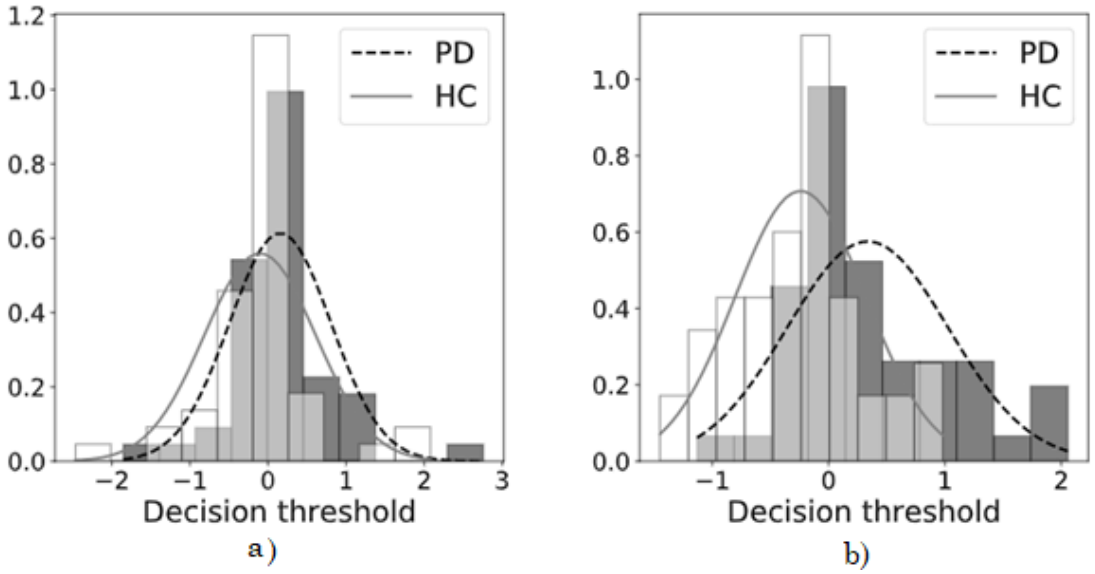

Fig. 3. Probability density distributions and histograms of the best scores in the SVM of PD patients vs. HC speakers without sex classification. a) Original vector. b) Vector with feature selection through Kruskal-Wallis test. Source: Created by the authors.

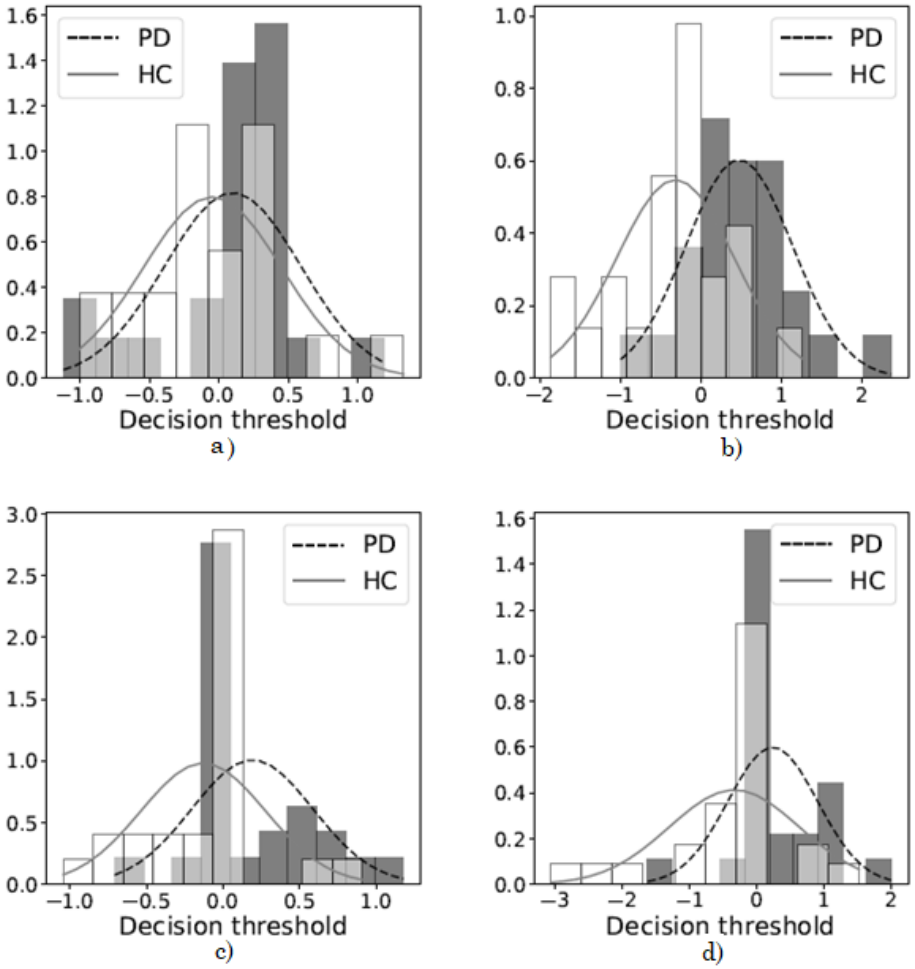

Fig. 4. Probability density distributions and histograms of the best scores in the SVM of a) male PD patients vs. male HC speakers with original vector, b) vector with feature selection through Kruskal-Wallis test, c) female PD patients vs. female HC speakers with original vector, and d) vector with feature selection through Kruskal-Wallis test. Source: Created by the authors.

Figure 5 shows an additional comparison of the best results obtained in the classification of PD patients vs. HC speakers considering the two experiments (with and without sex-based grouping).

The Receiver Operating Characteristic (ROC) curve represents the results in a more compact way and is a standard measure of performance in medical applications [20]. Fig. 5.B shows that the best results were obtained for female speakers. Also note that the feature selection process improves the results in all cases.

3.2 Estimation of speakers' dysarthria level: Multi-class classification and Regression.

To predict the severity of the dysarthria of patients with PD, we only used the group of features with the best performance in the bi-class classification.

The MDS-UPDRS-III scale evaluates the motor skills of different limbs (e.g., hands and arms). However, only one out of the 33 items in the scale is about speech assessment. This causes a limitation to the evaluation of patients' neurological state considering only speech recordings.

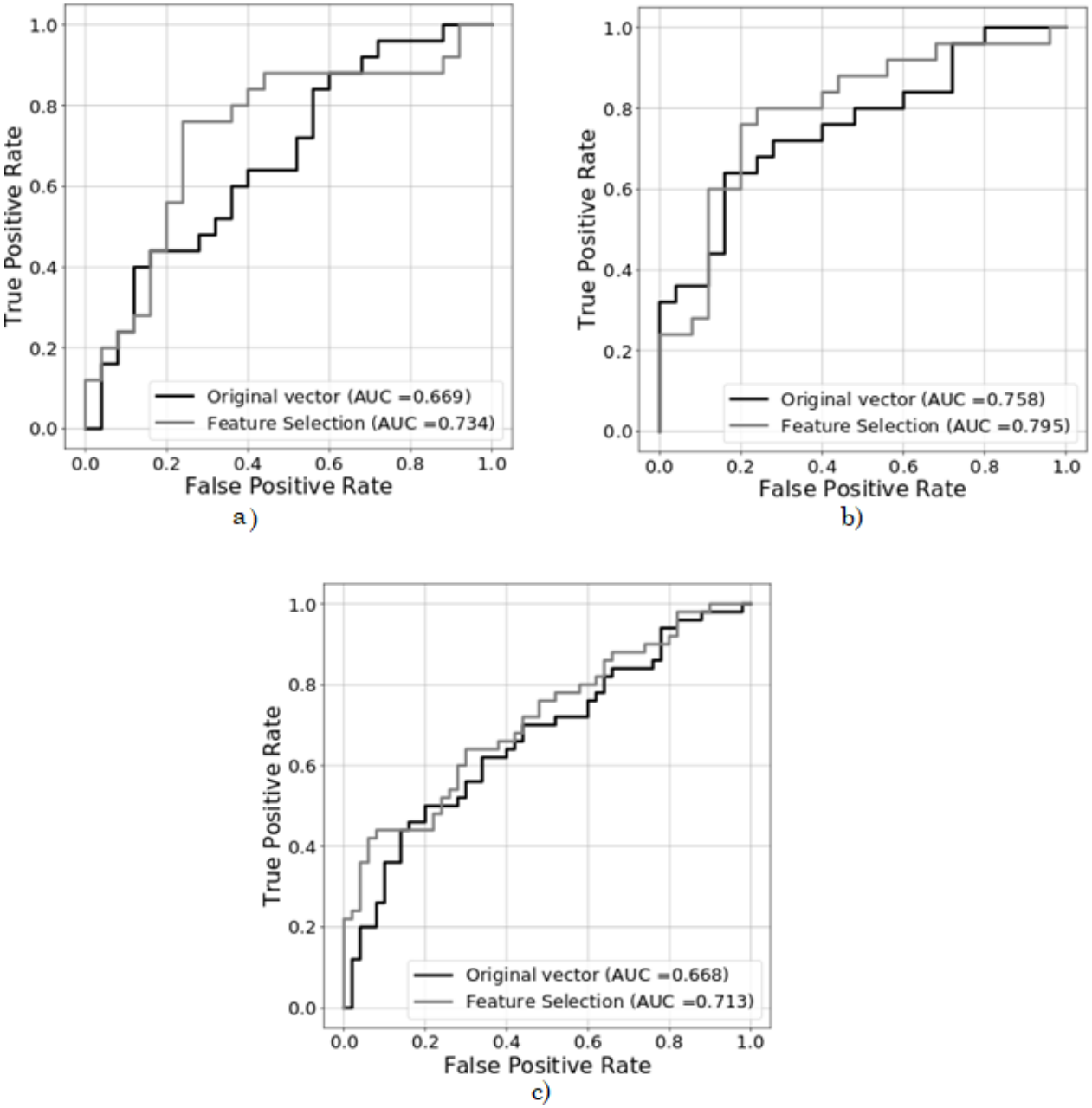

Fig. 5. ROC curves of the best results. a) Male PD vs Male HC. b) Female PD vs Female HC. c) PD vs HC without sex-based grouping. Source: Created by the authors.

The m-FDA scale was introduced in [13] with the aim of providing patients and clinicians with a tool that enables a more accurate evaluation of patients' dysarthria level. Such scale considers several aspects of speech, including breathiness, lip movement, palate movement, laryngeal capacity, tongue posture and movement, monotonicity, and intelligibility [13].

The m-FDA scale classifies speakers into four groups: speakers with m-FDA scores between 0 and 9 (N1), between 10 and 19 (N2), between 20 and 29 (N3), and over 30 (N4).

The results of the multi-class classification (N1 vs N2 vs N3 vs N4) are presented in Table 5, along with the accuracy, F1-score, Unweighted Average Recall, and confusion matrix. Note that, as in the bi-class classification experiments, the best results here were obtained when the female speakers were considered separately (51 % maximum accuracy), which is significantly higher than when men were studied separately (32 % accuracy) or when men and women were analyzed together (41 % accuracy).

Table 6 reports the results of the regression experiments in terms of m-FDA scores. It presents a maximum Spearman's rank correlation coefficient of 0.446, which was obtained for the group of female speakers. This low value is due to the fact that the tasks only considered vowel pronunciations. We believe that these results might improve if longer tasks with continuous speech signals are employed.

4. CONCLUSIONS

This paper introduces a novel method based on the HHT, and it shows that articulation-based features are suitable to discriminate between PD patients and HC speakers considering sustained modulated vowels. The feature selection strategy proposed in this paper, based on the results of Kruskal-Wallis tests, seems to improve the accuracy of the proposed approach. Additionally, a gender-based pre-grouping of speakers can enhance the results, especially in the case of female participants.

The main advantage of this approach is that only modulated vowels are considered to perform the evaluation. Other studies in the literature require longer tasks like reading texts and monologues, which are more expensive, time consuming, and even more invasive. Although better results were obtained in [5] with a similar approach, our methodology is focused on modeling the first two vocal formants using the HHT.

Table 5. Confusion matrix with the results of the classification of healthy controls and patients with PD at different stages of the disease. Source: Created by the authors.

	Male and female together				**Male**				**Female**			
	Acc=0.41, F1=0.32, UAR=0.36				Acc=0.32, F1=0.25, UAR=0.31				Acc=0.51, F1=0.46, UAR=0.45			
	N1	N2	N3	N4	N1	N2	N3	N4	N1	N2	N3	N4
N1	**23**	0	7	1	**5**	3	2	4	**13**	1	2	1
N2	10	**0**	10	1	8	**2**	1	1	2	**1**	4	2
N3	12	0	**14**	2	6	0	**4**	2	2	3	**9**	2
N4	5	0	11	**4**	1	3	2	**6**	1	2	3	**2**

Acc: Accuracy. **F1:** F1-score. **UAR:** Unweighted Average Recall
N1: Speakers with m-FDA scores between 0 and 9. N2: Speakers with m-FDA scores between 10 and 19
N3: Speakers with m-FDA scores between 20 and 29. N4: Speakers with m-FDA scores over 30.

Table 6. Results of the regression taking into account the different levels of the disease Source: Created by the authors.

Feature	**Male and female Together**		**Male**		**Female**	
	C_s	MAE	C_s	MAE	C_s	MAE
G1	**0.264**	11.032	0.196	11.502	0.228	10.840
G2	0.229	10.952	0.203	11.319	0.283	10.379
G3	0.244	10.575	**0.216**	11.377	**0.446**	9.234

C_s**:** Spearman rank-order correlation coefficient. **MAE:** Mean absolute error.

To the best of our knowledge, this is the first study that considers the HHT to model the temporal dynamics of F1 and F2.

The main motivation behind the use of the HHT to model the vocal formants is that it allows the analysis of the articulatory capacity of speakers, which, in general, is compromised in patients with PD. The results indicate that the HHT enables us to obtain relevant information in certain frequency bands that could be considered a suitable bio-marker to model the speech of PD patients.

Future work will consider nonlinear dynamical features to assess the complementarity between the information of HHT-based models and NLD features. Moreover, higher accuracies can be achieved considering measurements such as jitter, shimmer, noise measurements, periodicity, and stability of the first and second formant as well as the original signal.

5. ACKNOWLEDGEMENTS

The authors acknowledge to the GITA research group of the faculty of Engineering of the University of Antioquia. Also acknowledge to the Training Network on Automatic Processing of PAthological Speech (TAPAS) funded by the Horizon 2020 programme of the European Commission. Tomás Arias Vergara is under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS. This work was also funded by CODI from the University of Antioquia, grant number PRG2017-15530.

6. REFERENCES

[1] S. Anand and C. E. Stepp, "Listener Perception of Monopitch, Naturalness, and Intelligibility for Speakers With Parkinson's Disease," *J. Speech, Lang. Hear. Res.*, vol. 58, no. 4, pp. 1134–1144, Aug. 2015. http://pubs.asha.org/doi/10.1044/2015_JSLHR-S-14-0243

[2] S. Fahn, "Description of Parkinson's Disease as a Clinical Syndrome," *Ann. N. Y. Acad. Sci.*, vol. 991, no. 1, pp. 1–14, no. 991, pp. 1-14, Jun. 2003. https://doi.org/10.1111/j.1749-6632.2003.tb07458.x

[3] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and Cooccurrence of Vocal Tract Dysfunctions in the Speech of a Large Sample of Parkinson Patients," *J. Speech Hear. Disord.*, vol. 43, no. 1, pp. 47–57, Feb. 1978. https://doi.org/10.1044/jshd.4301.47

[4] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward Phonetic Intelligibility Testing in Dysarthria," *J. Speech Hear. Disord.*, vol. 54, no. 4, pp. 482–499, Nov. 1989. https://doi.org/10.1044/jshd.5404.482

[5] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Nöth, "Automatic Detection of Parkinson's Disease Based on Modulated Vowels," in *proc Interspeech*, San francisco, 2016, pp. 1190–1194. https://doi.org/10.21437/Interspeech.2016-1062

[6] J. Rusz *et al.*, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2171–2181, Aug. 2013. https://doi.org/10.1121/1.4816541

[7] S. Skodda, W. Visser, and U. Schlegel, "Vowel Articulation in Parkinson's Disease," *J. Voice*, vol. 25, no. 4, pp. 467–472, Jul. 2011. https://doi.org/10.1016/j.jvoice.2010.01.009

[8] J. R. Orozco-Arroyave *et al.*, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digit. Signal Process.*, vol. 77, pp. 207–221, Jun. 2018. https://doi.org/10.1016/j.dsp.2017.07.004

[9] R. R. Zhang, S. Ma, and S. Hartzell, "Signatures of the Seismic Source in EMD-Based Characterization of the 1994 Northridge, California, Earthquake Recordings," *Bull. Seismol. Soc. Am.*, vol. 93, no. 1, pp. 501–518, Feb. 2003. https://doi.org/10.1785/0120010285

[10] N. E. Huang and Z. Wu, "A review on Hilbert-Huang transform: Method and its applications to geophysical studies," *Rev. Geophys.*, vol. 46, no. 2, pp. 1-23, Jun. 2008. https://doi.org/10.1029/2007RG000228

[11] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease.," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014, pp. 342–347. Available: URL

[12] C. G. Goetz *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Mov. Disord.*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008. https://doi.org/10.1002/mds.22340

[13] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease," *J. Commun. Disord.*, vol. 76, pp. 21–36, Nov. 2018. https://doi.org/10.1016/j.jcomdis.2018.08.002

[14] N. E. Huang *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. London. Ser. A Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998. https://doi.org/10.1098/rspa.1998.0193

[15] J. C. Catford, "A practical introduction to phonetics", ed. Second, Oxford: Clarendon Press, 1998. Available: URL

[16] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "Spectral and cepstral analyses for Parkinson's disease detection in Spanish vowels and words," *Expert Syst.*, vol. 32, no. 6, pp. 688–697, Dec. 2015. https://doi.org/10.1111/exsy.12106

[17] L. R. Rabiner and R. W. Schafer, "Introduction to Digital Speech Processing," *Found. Trends® Signal Process.*, vol. 1, no. 1–2, pp. 1–194, Dec. 2007. Available: URL

[18] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS)," *J. Voice*, vol. 10, no. 1, pp. 59–66, Jan. 1996. https://doi.org/10.1016/S0892-1997(96)80019-1

[19] I. Hertrich and H. Ackermann, "Gender-Specific Vocal Dysfunctions in Parkinson's Disease: Electroglottographic and Acoustic Analyses," *Ann. Otol. Rhinol. Laryngol.*, vol. 104, no. 3, pp. 197–202, Mar. 1995. https://doi.org/10.1177/000348949510400304

[20] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomed. Signal Process. Control*, vol. 1, no. 2, pp. 120–128, Apr. 2006. https://doi.org/10.1016/j.bspc.2006.06.003