Cogn Comput
DOI 10.1007/s12559-017-9497-x

Parkinson's Disease and Aging: Analysis of Their Effect in Phonation and Articulation of Speech

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

Received: 10 October 2016 / Accepted: 18 July 2017

Abstract Parkinson's disease (PD) is a neurological disorder that affects the communication ability of patients. There is interest in the research community to study acoustic measures that provide objective information to model PD speech. Although there are several studies in the literature that consider different characteristics of Parkinson's speech like phonation and articulation, there are no studies including the aging process as another possible source of impairments in speech. The aim of this work is to analyze the vowel articulation and phonation of Parkinson's patients compared with respect to two groups of healthy people: (1) young speakers with ages ranging from 22 to 50 years and (2) people with ages matched with respect to the Parkinson's patients. Each participant repeated the sustained phonation of the five Spanish vowels three times and those utterances per speaker are modeled by using phonation and articulation features. Feature selection is applied to eliminate redundant information in the features space, and the automatic discrimination of the three groups of speakers is performed using a multi-class Support Vector Machine (SVM) following a one vs. all strategy, speaker independent. The results are compared to those obtained using a cognitive-inspired classifier which is based on neural networks (NN). The results indicate that the phonation and articulation capabilities of young speakers clearly differ from those exhibited by the elderly speakers (with and without PD). To the best of our knowledge, this is the first paper introducing experimental evidence to support the fact that age matching is necessary to perform more accurate and robust evaluations of pathological speech signals, especially considering diseases suffered by elderly people, like Parkinson's. Additionally, the comparison among groups of speakers at different ages is necessary in order to understand the natural change in speech due to the aging process.

Keywords Parkinson's disease · Phonation · Articulation · Aging voice · Multi-class SVM · Neural networks

Introduction

There exist different physiological changes in people's life due to several reasons including aging and disease conditions. There are changes in speech that result from the natural aging process [1]; however, when those disturbances appear due to a disease, the changes must be analyzed with detail in order to state which should be the treatment required to ameliorate the state of the patient. As the speech of elderly people can change due to the aging process or due to the presence of a disease (or both), the description and classification of features in speech that reflect such differences is a topic that deserves special attention. Since Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder worldwide and affects about 2% of people older than 65 years [2], this study addresses the analysis of phonation and articulation characteristics of speech of people with PD and compare those features with respect

✉ T. Arias-Vergara
tomas.arias@udea.edu.co

J. C. Vásquez-Correa
jcamilo.vasquez@udea.edu.co

J. R. Orozco-Arroyave
rafael.orozco@udea.edu.co

[1] Faculty of Engineering, Universidad de Antioquia, 50010 Medellín, Colombia

[2] Pattern Recognition Laboratory, Friedrich Alexander Universitaï Erlangen-Nuïrnberg, 91058 Erlangen, Germany

Published online: 04 August 2017

to two groups of speakers: young and elderly people, both with normal and healthy physical and mental conditions.

There are motor and non-motor symptoms associated with PD and the majority of patients exhibit voice and speech impairments due to the disease [3]. Additionally, the changes in organs and tissues involved in voice production which are associated with the aging process include facial skeleton growth [4], pharyngeal muscle atrophy [5], tooth loss [6], reduced mobility of the jaw [7], and tongue musculature atrophy. These changes alter the phonation and articulation dimensions of speech, for instance elderly people exhibit a significantly greater frequency perturbation than that in young speakers [8] and there are also differences in the stability of the frequency and amplitude of vocal fold vibration relative to young and middle-aged adults [9]. A reduction in the frequency of the first three vocal formants has also been observed [10]. Regarding the speech impairments of PD patients, several dimensions of speech are affected including phonation, articulation, prosody, and intelligibility [11, 12]. Phonation impairments in PD patients include inadequate closing of the vocal fold and vocal fold bowing [13], which generates stability and periodicity problems in vocal fold vibration [14]. The articulation problems are mainly related with reduced amplitude and velocity of lip, tongue, and jaw movements [15], generating a reduced articulatory capability in PD patients to produce vowels [16] and to produce continuous speech [17]. These deficits reduce the communication ability of PD patients and make their normal interaction with other people difficult.

There are many contributions in the literature analyzing the impact of PD in the articulation and phonation capability of the patients. In [16], the authors compare the speech of 68 PD patients and 32 age-matched healthy controls (HC). The vowels /a/, /i/, and /u/ were extracted from a text which was read by the speakers. The values of the first two formants (F_1 and F_2) are calculated from each vowel to form the vowel space, i.e., F_1 vs F_2. The vowel articulation is analyzed with the triangular Vowel Space Area (tVSA), and the Vowel Articulation Index (VAI). The authors conclude that VAI is reduced in PD speakers compared with respect to the HC group. In [18], speech recordings of 38 PD patients and 14 HC are analyzed. The participants repeated three sentences several times. The vowels /a/, /i/, and /u/ are extracted from the recordings and several articulation features are estimated including tVSA, natural logarithm of tVSA, Formant Centralization Ratio (FCR), and the ratio F_{2i}/F_{2u}, where F_{2i} and F_{2u} are the values of the second formants extracted from the vowels /i/ and /u/, respectively. The results indicate that FCR and F_{2i}/F_{2u} are highly correlated ($r = -0.90$); additionally, the authors conclude that with both measures it is possible to differentiate PD patients from HC speakers. In [19], the authors performed vowel articulation analyses in recordings of 20 early PD patients and 15 aged-matched HC. The speech tasks considered in this study include sustained phonations of the Czech vowel /i/, repetition of short sentences, reading of a text with 80 words, and a monologue of approximately 90 sec duration. The articulation analysis was performed with different acoustic measures such as tVSA, VAI, F_1 and F_2, and the ratio F_{2i}/F_{2u}. The monologue was the most suitable task to differentiate speech of early PD patients and HC speakers, with classification accuracies of up to 80%. The authors claim that, based on their results, sustained phonation may not be suitable to evaluate vowel articulation in early PD; however, this assertment contradicts other studies in the state of the art indicating that the analysis of sustained phonations seems to be a good alternative to assess Parkinson's speech [14, 20–23]. Besides the articulation analysis, several studies consider phonation in speech of people with PD. In [20], phonation features are calculated upon sustained phonations of the English vowel /a/. The database for those experiments includes 263 phonations performed by 43 subjects (33 PD patients and 10 HC). A total of 132 measures are considered including different variants of jitter and shimmer, several noise measures, Mel Frequency Cepstral Coefficients (MFCCs), and nonlinear measures. Two different classification strategies are compared, random forest (RF) and Support Vector Machines (SVM) whit Gaussian kernel. The classifiers are trained following a 10-fold cross validation strategy, i.e., the 263 phonations are split into two subsets: training, which consists of 90% of the data (237 phonations), and test subset, which consists of the remaining 10% of the data (26 phonations). The process is repeated 100 times, randomly permuting the train and test subsets. The authors report accuracies of up to 98.6% using 10 dysphonia features; however, the speaker independence is not satisfied. Note that the database contains 263 phonations from 43 subjects, which means that each speaker repeated the phonation about 6 times, but the authors did not assure that all of the repetitions were in the same subset (train or test). This strategy leads to methodological issues because the recordings are mixed into the train and test subsets, producing optimistic results and possible biased conclusions. In [24], phonation and articulation analyses are performed considering recordings of sustained vowels performed by a total of 100 speakers. The five Spanish vowels are uttered three times by 50 PD patients and 50 age-matched HC. Articulation analysis is performed with different acoustic measures such as F_1 and F_2, tVSA, and VAI. Additionally, three new measures are introduced: the vocal prism volume, the Vowel Pentagon Area (VPA), and the vocal polyhedron. Phonation is evaluated trough a set of measures that includes jitter, shimmer, and the correlation dimension (D2). The authors

performed the automatic classification of PD speakers and HC, and report accuracies of 81% when phonation and articulation features are combined. Although each speaker repeated the phonations several times, the authors report that the speaker independence is satisfied, i.e., the three repetitions of the same speaker are in the train or test subsets but not mixed. Besides the analysis of phonation features to detect/discriminate Parkinson's disease, there are other works focused on the understanding of several diseases that negatively impact speech. For instance in [25] the authors present an analysis of the neural pathways involved in the production of phonation and perform experiments to show their connection to different phenomena like vocal fold stiffness which is present in most of the Parkinson's patients. Additionally, in [14] several diseases are considered (Parkinson's, cleft lip and palate, and laryngeal cancer) and analyzed by modeling sustained phonations of vowels. According to the results, in order to obtain a more accurate description of each disorder, it is necessary to consider different features, for instance phonation features are more affected in patients with laryngeal cancer than in patients with cleft lip and palate.

Regarding the studies analyzing the impact of aging in speech, in [9] the authors consider sustained phonations of the English vowel /a/ and compute fifteen phonation measures of the Multi-Dimensional Voice Program (MDVP) model 4305. The set of measures includes F_0, jitter, Pitch Perturbation Quotient (PPQ), Relative Average Perturbation (RAP), variability of F_0, Amplitude Perturbation Quotient (APQ), shimmer, Noise to Harmonics Ratio (NHR), and others. A total of 44 speakers (21 male and 23 female) aged between 70 and 80 years were considered and compared with respect to the norms for young and middle-aged adults published in [26]. The authors perform statistical analyses and report that the voice of elderly people is significantly different (usually poorer) than the voice of young and middle-aged adults. In [27], the authors calculate several phonation measures to assess the stability of vocal fold vibration and to quantify the noise in the voice of 159 younger speakers with ages between 18 and 28 years, and 133 older adults with ages between 63 and 86 years. The authors conclude that the instability of the vocal fold vibration increases with age. The Dysphonia Severity Index (DSI) was also measured and only older females exhibited higher values than those in younger females. No statistical differences were observed between younger and older males. Other study that evaluates the influence of aging in the speech of elderly people considering phonation and articulation analyses is presented in [28]. A total of 27 young speakers with mean age of 25.6 years and 59 older people with mean age of 75.2 years is considered. Each participant was asked to read a set with 22 consonant-vowel-consonant (CVC) words. The vowels and oral stops of each word where extracted and analyzed using Praat [29]. The authors analyze several acoustic properties including F_0, the first three formants and the Voice Onset Time (VOT). F_0 allows them to study possible changes in the fundamental frequency of vocal fold vibration, and the first three formants give information about the position of the tongue (forward, backward, or closer to the palate), and the VOT provides information about the timing to produce the oral stops. According with the results, there is a clear lowering of F_0 with age for women, and a raising of F_0 with age for men. This finding is consistent with previous reports such as [8]. The authors highlight also that older men showed shorter VOTs than both younger men and younger women, which is also reported in [30]. A greater variability in F_0, the three formants, and the VOT is systematically observed in the speech productions by older adults compared to their younger same-sex counterparts. As the natural aging process in humans carries several alterations in speech production and perception, the impact of aging in the detection of voice disorders is still an open problem and its relevance in the clinical practice was recently studied in [31].

Additionally, there are several works in the state-of-the-art where cognitive-inspired systems are proposed to model speech. For instance in [32] the authors present a system based on multi-scale product with fuzzy logic to separate voiced and unvoiced segments in speech signals. Additionally, a comb filter is applied to reduce noise in the voiced segments while the classical spectral subtraction is applied upon the unvoiced frames. According to the results, the cognitive-based approach outperforms other state-of-the-art methods to reduce noise in speech signals recorded in non-controlled acoustic conditions. In [33], the authors perform the automatic detection of affective states from speech. They compared a classical model based on Gaussian Mixture Models (GMM) with a cognitive inspired multi-layer perceptron (MLP). Several feature sets typically used in speech processing such as MFCCs, energy content, pitch and others are used. According to their results, the GMM-based approach is more suitable than the MLP to model emotional speech signals. Also in [34] the authors present a special issue with several contributions considering cognitive systems to model different phenomena of speech.

Considering the increasing relevance of cognitive systems to model speech signals, the proposed approach is compared to a cognitive-inspired classifier which is based on a multi-class neural network. According to our results, the cognitive-inspired classifier is a good alternative for the multi-class task of discriminating Parkinson's patients, elderly healthy speakers and young healthy speakers. Additionally, the reviewed state-of-the-art shows that most of the

studies are focused on comparing Parkinson's speech with respect to the speech of age- and gender-matched healthy controls. However, abnormal vocal fold vibration and articulatory problems may appear in healthy speakers due to the aging process. Thus, the age is a confounding factor when automatic systems are used for diagnosis. The aim of this paper is to evaluate the effect of Parkinson's disease and aging in the phonation and articulation processes of speech.

The rest of the paper is organized as follows: "Data Description" includes the description of the data, "Methodology" includes details of the methodology presented in the paper. "Feature Extraction" describes the features computed to model the speech signals, "Experiments and Results" describes the experiments and results, "Cognitive-Inspired Classifier" introduces a cognitive-inspired multi-class classifier and includes the obtained results to be compared with respect to those obtained with the proposed approach, and finally "Conclusions" includes the conclusions derived from this work.

Data Description

Three groups of speakers will be compared in this paper: 50 patients with PD, 50 age and gender -matched healthy controls (aHC), and 50 healthy young speakers (yHC). Each group contains 25 male and 25 female. The participants are Spanish native speakers and were asked to pronounce the five Spanish vowels in a sustained manner. The age of PD patients ranges from 33 to 81 (mean 61.14 ± 9.61), the age of the aHC group ranges from 31 to 86 (mean 60.9 ± 9.46), and the age of the yHC group ranges from 17 to 52 (mean 22.94 ± 6.06). The recordings were captured in a sound-proof booth using a professional audio-card and

Table 1 Detailed information of the PD patients and healthy speakers

M-PD			M-aHC	M-yHC	W-PD			W-aHC	W-yHC
AGE	UPDRS-III	t	AGE	AGE	AGE	UPDRS-III	t	AGE	AGE
81	5	12	86	52	75	52	3	76	38
77	92	15	76	32	73	38	4	75	34
75	13	1	71	30	72	19	2.5	73	27
75	75	16	68	28	70	23	12	68	24
74	40	12	68	26	69	19	12	65	24
69	40	5	67	26	66	28	4	65	23
68	14	1	67	26	66	28	4	64	23
68	67	20	67	26	65	54	8	63	23
68	65	8	67	24	64	40	3	63	22
67	28	4	65	23	62	42	12	63	22
65	32	12	64	23	61	21	4	63	22
65	53	19	63	22	60	29	7	62	21
64	28	3	63	22	59	40	14	62	21
64	45	3	62	22	59	71	17	61	21
60	44	10	60	22	58	57	1	61	21
59	6	8	59	21	57	41	37	61	21
57	20	0.4	56	21	57	61	17	60	19
56	30	14	55	20	55	30	12	58	19
54	15	4	55	20	55	43	12	57	19
50	53	7	54	20	55	30	12	57	19
50	19	17	51	19	55	29	43	55	18
48	9	12	50	18	54	30	7	55	18
47	33	2	42	18	51	38	41	50	18
45	21	7	42	18	51	23	10	50	17
33	51	9	31	17	49	53	16	49	17

t time post PD diagnosis in years, *M-PD* men Parkinson's disease, *M-aHC* men age-matched healthy controls, *M-yHC* men young healthy controls, *W-PD* women Parkinson's disease, *W-aHC* women age-matched healthy controls, *W-yHC* women young healthy controls, *MDS-UPDRS* Movement Disorder Society-Unified Parkinson's Disease Rating Scale

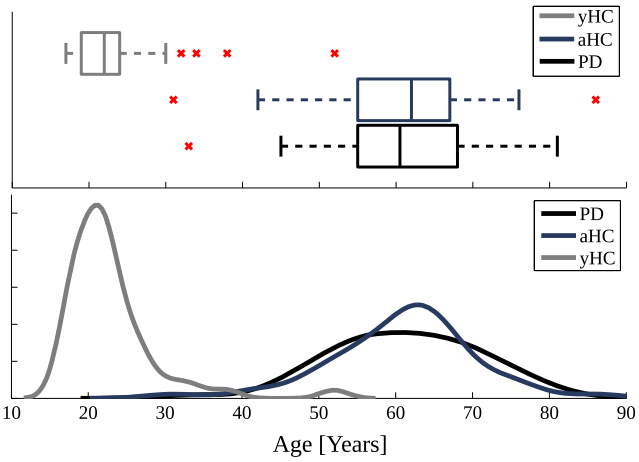


Fig. 1 Age distribution of the PD patients (*black curve*), aHC (*dark-blue curve*), and yHC (*light-gray curve*) groups

a dynamic omni-directional microphone. The speech signals were sampled at 44.1 kHz with 16-bit resolution. All of the PD patients were diagnosed by a neurologist expert and were labeled according to the motor sub-scale of the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS-III) [35]. The patients were in ON-state during the recording session, i.e., no more than 3 h after the morning medication. None of the speakers in the healthy groups had symptoms associated with PD or any other neurological disease.

Table 1 displays details of the age, MDS-UPDRS-III scores, and the time after the PD diagnosis. Male and female are presented separately. For the aHC and yHC groups, only the age values are provided.

Figure 1 shows the age distribution from the three groups of speakers represented with box plots (top figure) and fitted kernel densities (bottom figure). It can be observed that there are 4 outliers in the yHC group, two in the PD and one in the aHC. As the construction of this database started with the PD patients and the original group included one young patient (33 years) and one old patient (81 years), the outliers of the other two groups were included to compensate the unbalance introduced in the PD group.

Methodology

Figure 2 illustrates the methodology proposed in this study. It comprises four main stages. (1) Recording and preprocessing of the five Spanish vowels uttered by the participants. (2) Computation of the features upon the voice signals (the five Spanish vowels are considered per speaker) in order to model the articulation and phonation dimensions, forming two feature matrices $[\Psi_{Pho}]_{m\times n_{Pho}}$ and $[\Psi_{Art}]_{m\times n_{Art}}$ for phonation and articulation models, respectively. The features extracted form the five Spanish vowels are considered together in all of the experiments. m is the number of speakers, n_{Pho} is the number of phonation features, and n_{Art} is the number of articulation features. (3) Feature selection and relevance analysis is performed by using principal component analysis (PCA). In this stage, the feature space is reduced, thus the new feature matrices are $[\hat{\Psi}_{Pho}]_{m\times\rho_{Pho}}$ and $[\Psi_{Art}]_{m\times\rho_{Art}}$, where $\rho_{Pho} < m$ and $\rho_{Art} < m$. (4) The automatic discrimination of the three groups of speakers (PD, aHC, and yHC) is performed by using two different multi-class classifiers, one is based on SVM and the other one is based on NN. More details of each stage are presented in the following subsections.

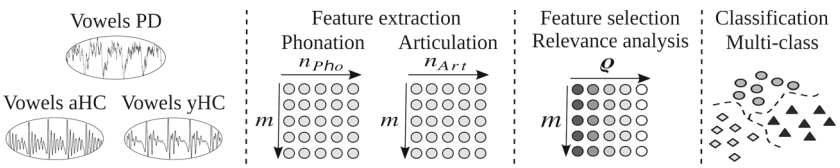
Voice Recording and Pre-processing

The voice signals are recorded in a sound-proof booth, using a professional audio card (M-Audio, ref. Fast Track Pro.) and an omni-directional microphone (Shure, ref. SM63) connected using professional cabling. All the recordings are normalized in amplitude between -1 and $+1$. Although the acoustic conditions are quite controlled in our recordings, a cepstral mean subtraction procedure is applied in order to remove possible bias introduced by changes in the distance to the microphone during the recording session and among speakers [36].

Feature Extraction

The recordings of the five Spanish vowels uttered in a sustained manner are modeled considering phonation and articulation measures. Phonation features evaluate disorders in the vocal folds vibration, and articulatory features (extracted from sustained phonations) evaluate changes in the position of the tongue while different vowels are produced. Each feature is calculated on a frame basis. The length of each frame and the corresponding overlap depends on the nature of the feature, i.e., there are long-term or short-term analyses. Four functionals are computed per feature: mean, standard deviation, kurtosis, and skewness. Details

Fig. 2 Methodology



of the computed features are presented in the following subsections.

Phonation Measures

- *Jitter and shimmer:* Variations in the frequency and amplitude of the pitch period are defined as jitter and shimmer, respectively.
- *Amplitude Perturbation Quotient (%):* This feature measures the long-term variability of the peak-to-peak amplitude of the pitch period with a smoothing factor of 11 periods [37].
- *Pitch Perturbation Quotient (%):* This feature measures the long-term variability of the fundamental period (pitch) with a smoothing factor of 5 periods [37].

  The jitter, shimmer, APQ, and PPQ are used to model the stability of vocal fold vibration. Additionally, several noise features are extracted with the aim of modeling glottal and turbulent noise that appears due to the abnormal closing of the vocal fold which is typically observed in people with loss of control of the vocal fold, like PD patients.
- *Harmonics to Noise Ratio (*HNR*):* The computation of HNR is based on the assumption that a sustained phonation has two components: a quasi-periodic component that is the same from cycle-to-cycle and a noise component that has a zero mean amplitude distribution. HNR is determined as the relation between the acoustic energy of the average harmonic structure and the noise component of the voice signal. HNR is calculated using the method presented in [38].
- *Cepstral Harmonics to Noise Ratio (*CHNR*):* This measure is based on the method presented in [39] for the calculation of the HNR in the cepstral domain.
- *Normalized Noise Energy (*NNE*):* This is an acoustic measure introduced in [40] to evaluate the noise components of pathological voices.
- *Glottal to Noise Excitation Ratio (*GNE*):* This measure was introduced in [41] to determine whether a voice signal is generated from vocal fold vibration or from turbulent noise originated in the vocal tract.

Articulation Measures

- *Vocal formants:* The formants are defined as acoustic energy accumulated in certain frequency bands. The energy is generated from the shape and position formed by the articulatory organs involve in the speech production process. Commonly, the F_1 and and F_2 are used to measure articulatory impairments during sustained phonation. Additionally, F_1 and and F_2 are used for the estimation of tVSA, VPA, and FCR.

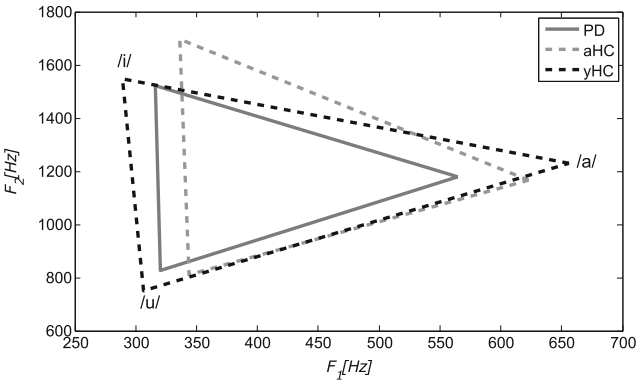


Fig. 3 Vowel triangles for the PD patients (*dark-gray solid triangle*), aHC (*gray dotted triangle*), and yHC groups (*black dotted triangle*)

- *Triangular Vowel Space Area (*tVSA*):* This measure is used to model possible reduction in the articulatory capability of speakers. Such a reduction is observed as a compression of the area of the vocal triangle, i.e., a reduced value of tVSA. The main hypothesis is that young speakers have a better articulation capability than elderly speakers (either healthy or with PD), thus they are able to move their tongue with greater amplitudes and they are able to hold it longer in certain positions according to the pronounced phonation. Figure 3 displays the average vocal triangles obtained considering phonations of the PD group (solid dark-gray lines), aHC group (dotted gray lines), and yHC group (dotted black lines). Note that PD patients exhibit a compressed tVSA compared to those obtained with the yHC and aHC groups. The largest triangle of the young group confirms the hypothesis and indicates that they have a better articulation capability.
- *Vowel Pentagon Area (*VPA*):* This measure allows the quantification of articulatory movements performed when producing the five Spanish vowels. This measure was introduced in [24] to evaluate articulatory deficits of people with Parkinson's disease. Figure 4 shows the

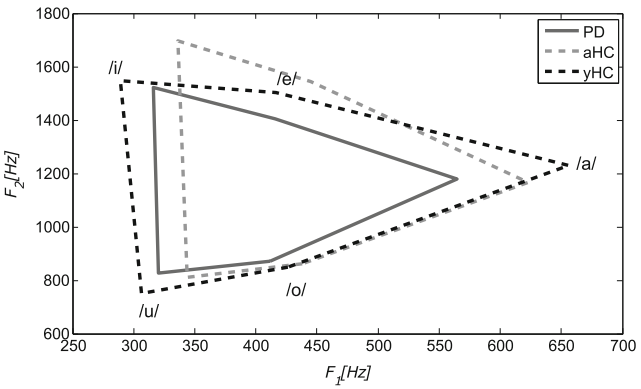


Fig. 4 Vowel pentagon for the PD patients (*dark-gray solid polygon*), aHC (*gray dotted polygon*), and yHC groups(*black dotted polygon*)

vocal pentagon obtained with phonations of the PD patients, aHC, and yHC. The largest VPA is obtained with phonations of the yHC group, which confirms the result obtained with tVSA.

- *Formant Centralization Ratio* (FCR): This measure was introduced by Sapir et al. in [18] to analyze changes in the vocal formants with a reduced inter-speaker variability, it can be used to improve the discrimination of people with PD and healthy speakers.
- *Mel Frequency Cepstral Coefficients*: These coefficients are a smoothed representation of the speech spectrum considering information of the scale of the human hearing. They are widely used to model articulatory problems in the vocal tract [42]. In this study, 12 MFCCs along with their first- and second-order derivatives are considered. The derivatives are included to capture the dynamic information of the coefficients.

Feature Selection

A relevance analysis was performed for the combination of the five Spanish vowels using PCA with a modification that allows to obtain a reduced representation formed with the original descriptors rather than a transformed representation of the feature space. This approach was successfully used in previous studies where the reduction of redundancy and dimensionality of the feature space yield improved results [43]. PCA is based on the variance–maximization with the aim of find the ρ most relevant features of the original space $X \in \mathbb{R}^{m \times p}$ (n: number of observations, p: number of original features), which makes it possible to build the subspace representation $X_\rho \in \mathbb{R}^{m \times \rho}$, ($\rho < p$), where each of the ρ variables are not correlated with each other. Although PCA is commonly used as a reducing dimension technique, it can also be used for feature selection based on the relevance analysis of each feature, in such a way that a subset of the original feature space can be obtained [44]. The relevance of each feature in the original feature space can be identified according to $\boldsymbol{\varrho}$, which is defined according to Eq. 1, where λ_j and $\boldsymbol{v}_j$ are the eigenvalues and eigenvectors of the original feature matrix.

$$\boldsymbol{\varrho} = \sum_{j}^{\rho} \left| \lambda_j \boldsymbol{v}_j \right| \tag{1}$$

The values of $\boldsymbol{\varrho}$ for each feature are related to the contribution from each feature of the original space to each principal component. The original feature that is more correlated with each principal component will have the highest value of $\boldsymbol{\rho}$. In that way, the original feature can be recovered from the principal component and added to the feature subspace.

Data Distribution: Train, Development, and Test

The distribution of the data is performed into two stages. In the first stage, there are 50 speakers per group (PD, aHC, and yHC). Forty-five speakers of each group are considered to form the training subset and the remaining five speakers are considered to form the test subset. The second stage of the data distribution consists of dividing the subset of 45 train speakers into two sets: train and development. Forty of the 45 train speakers per group are considered to train the classification models and the remaining five speakers are considered to optimize the parameters of the classifiers, i.e., development set. Once the models are optimized, they are tested upon the five samples that were separated in the first stage of the data distribution. The first stage of the data distribution is performed 10 times to compute the confidence intervals of the classification accuracies. The second stage of the data distribution is repeated 9 times for a better optimization of the parameters in the classifier. This procedure is illustrated in Fig. 5. We are aware of the fact that this procedure is slightly optimistic; however, considering that only two parameters are optimized, the bias is minimal.

Classification

Two different strategies are performed to assess the influence of age in the voice of the three groups of speakers: PD patients, aHC, and yHC. The first strategy consists of training three SVMs with Gaussian kernel to perform three different classification experiments, respectively: (1) yHC vs. aHC groups, (2) PD vs. aHC, and (3) yHC vs PD. The other strategy consists of a multi-class SVM to automatically discriminate among the three groups of speakers: PD,

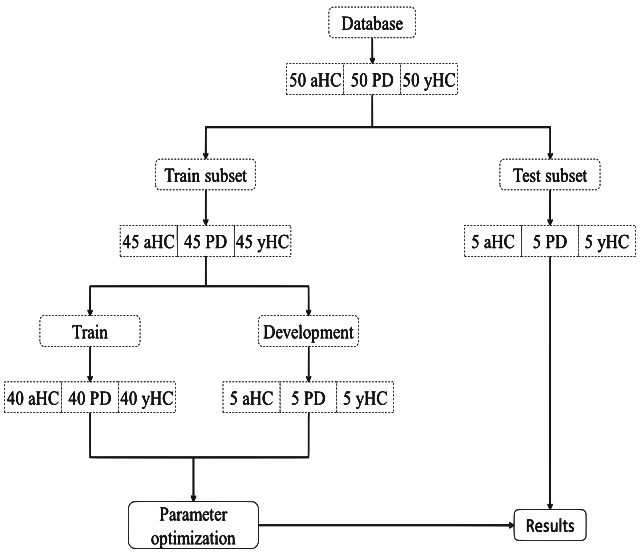


Fig. 5 Train, development, and test data distribution

aHC, and yHC. The aim of this strategy is to state to what extent the age is a confounding factor in situations where the system that discriminates between PD vs. HC includes young and/or age-matched healthy speakers in its training set. Further details about these two strategies are described in the following subsections.

Multi-class SVM It is necessary to introduce the case of a binary SVM classifier. The goal of an SVM is to discriminate data points by using a separating hyperplane which maximizes the margin between two classes. When some errors in the process of finding the optimal hyperplane are allowed, the classifier is known as a soft-margin SVM (SM-SVM) and the decision function is expressed as

$$t_k(\mathbf{w}^T\phi(\boldsymbol{x}_n)+b) \geq 1-\xi_n, n=1,2,3,...,N \tag{2}$$

where $t_k \in \{-1,+1\}$ are the class labels, $\phi(x)$ is the transformed feature space, $\mathbf{w}$ is the vector normal to the hyperplane, b is the bias parameter, N is the number of samples, and $\xi \geq 0$. In the SM-SVM approach, the errors in classification due to the overlapped classes are allowed. However, these errors are penalized using the slack variables ξ_n, which are introduced as the cost for misclassified data points. Figure 6 shows the influence of the slack variables in a SM-SVM. Considering class $y(x)=+1$ as reference, the slack variables take values of $\xi_n=0$ for each data point that lies on the margin or in the correct side of the margin (red circles). For the data points inside the margin and in the correct side of the decision boundary, the slack variables take values in the range of $0<\xi_n\leq 1$ (green circles). For those data points on the wrong side of the margin, the values of the slack variables are $\xi_n>1$ (blue circles) [45]. Now the goal is to maximize the margin while penalizing the data points for which $\xi_n>1$. Therefore, we wish to minimize

$$\text{minimize} \quad C\sum_{n=1}^{N}\xi_n+\frac{1}{2}\|\mathbf{w}\|^2 \tag{3}$$

where the parameter C controls the trade-off between ξ_n and the margin [45]. This is a convex optimization problem

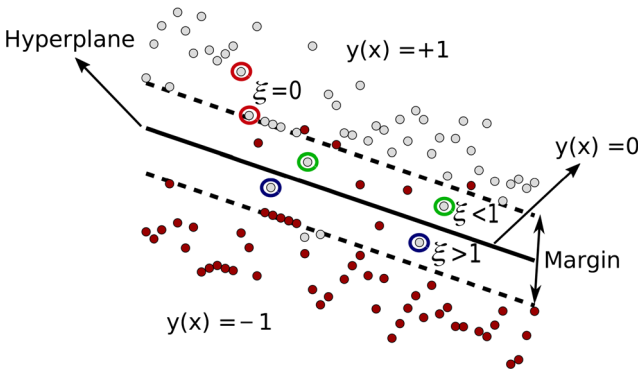


Fig. 6 Soft-margin SVM

where the goal is to minimize Eq. 3 subject to the constrains introduced in Eq. 2. One way to solve the problem is in its dual formulation using the Lagrange multipliers. The main idea in the dual formulation is to construct the Lagrange function from the primal function (objective function). The Lagrange function of the primal problem is expressed as

$$L=\frac{1}{2}\|w\|^2+C\sum_{n=1}^{N}\xi_n-\sum_{n=1}^{N}\alpha_n\{t_n(\mathbf{w}^T\phi(\boldsymbol{x}_n)+b)-1+\xi_n\}-\sum_{n=1}^{N}\mu_n\xi_n \tag{4}$$

where $\alpha_n \geq 0$ and $\mu_n \geq 0$ are Lagrange multipliers. In order to compute b the Karush-Kuhn-Tucker (KKT) conditions are verified. The set of KKT conditions are expressed as [46]:

1. Primal constrains

$$\alpha_n \geq 0 \tag{5}$$

$$t_n(\mathbf{w}^T\phi(\boldsymbol{x}_n)+b)-1+\xi_n \geq 0 \tag{6}$$

2. Complementary slackness

$$\alpha_n(t_n(\mathbf{w}^T\phi(\boldsymbol{x}_n)+b)-1+\xi_n)=0 \tag{7}$$

$$\mu_n\xi_n=0 \tag{8}$$

3. Dual constrains

$$\alpha_n \geq 0 \tag{9}$$

$$\mu_n \geq 0 \tag{10}$$

Now $\mathbf{w}$, b, and ξ_n have to vanish for optimality. This is accomplished estimating the partial derivatives of L with respect the primal variables:

$$\begin{aligned}\frac{\partial L}{\partial b} &= \sum_{n=1}^{N}\alpha_n t_n=0\\ \frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w}-\sum_{n=1}^{N}\alpha_n t_n\phi(\boldsymbol{x}_n)=0\\ \frac{\partial L}{\partial \xi_n} &= C-\alpha_n-\mu_n=0\end{aligned} \tag{11}$$

Now the dual Lagrangian formulation is expressed as

$$L_D=\sum_{n=1}^{N}\alpha_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m t_n t_m k(\boldsymbol{x}_n,\boldsymbol{x}_m) \tag{12}$$

Subject to

$$0\leq\alpha_n\leq C \tag{13}$$

$$\sum_{n=1}^{N}\alpha_n t_n=0 \tag{14}$$

where $k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}')$ is known as the kernel function. Data points where $\alpha_n > 0$ are called support vectors and must satisfy the condition

$$t_n(\mathbf{w}^T \phi(\boldsymbol{x}_n) + b) = 1 - \xi_n \tag{15}$$

From Eq. 11, it can be observed that if $\alpha_n < C$ then $\mu_n > 0$. It follows from Eq. 8 that $\xi_n = 0$, which indicates that such data points lie on the margin. The data points where $\alpha_n = C$ can lie inside the margin and in this case the slack variables can be either $\xi_n \leq 1$ or $\xi_n > 1$. The support vectors for which $0 < \alpha_n < C$ have $\xi_n = 0$. Substituting in Eq. 15, it follows that those support vectors will satisfy

$$t_n \left(\sum_{m \in \mathcal{S}} \alpha_m t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m) + b \right) = 1 \tag{16}$$

To compute b, a numerically stable solution is obtained by averaging.

$$b = \frac{1}{N_{\mathcal{M}}} \sum_{n \in \mathcal{M}} \left(t_n - \sum_{m \in \mathcal{S}} \alpha_m t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m) \right) \tag{17}$$

where $\mathcal{M}$ and $\mathcal{S}$ represent the set of data points such that $0 < \alpha_n < C$ and the set of total support vectors, respectively [45]. The SM-SVM described before corresponds to the case of overlapped data with a linear decision boundary. However, in many applications, a linear decision function may not exist or is not optimal to discriminate overlapped data. In those cases, kernel functions are considered to build a non-linear decision boundary. One of the most common kernel used in Pattern Recognition is the Gaussian kernel, which is expressed as

$$k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \mathrm{e}^{-\frac{2}{\gamma^2}\|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2} \tag{18}$$

where γ is the bandwidth of the Gaussian kernel. In this study, the parameters C and γ are optimized in a grid-search up to powers of ten with $1 \leq C \leq 10^4$ and $1 \leq \gamma \leq 10^3$ and the selection criterion is based on the highest accuracy obtained in the development subset. The automatic classification of the three classes is performed following a "one vs. all" strategy: three binary classifiers are considered, each classifier has a target class which is compared with respect to the combination of the remaining two classes, i.e., PD vs. aHC + yHC, aHC vs. PD + yHC, and yHC vs. PD + aHC. A total of three scores per recording are obtained. Recordings with maximum positive classification score are associated with the corresponding target class. If the maximum score is not positive, the recording should belong to one of the remaining two classes, thus a second binary classification is performed to decide in favor of one of those two remaining classes. Figure 7 shows a diagram that illustrates this strategy.

Classification and Class-Separability Analysis

Three different SVMs are trained. The first SVM is trained considering only the yHC and aHC groups. This experiment is performed to evaluate the discrimination capability of the proposed system when the age is the only difference between the two classes. In the second experiment, only

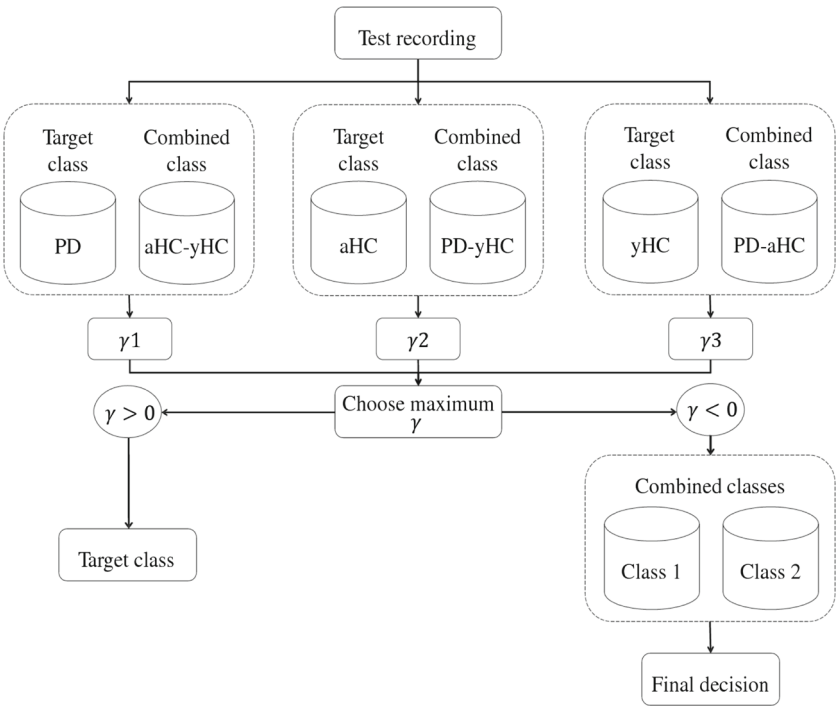


Fig. 7 "One vs all" strategy addressed to train/test a three-class SVM classifier

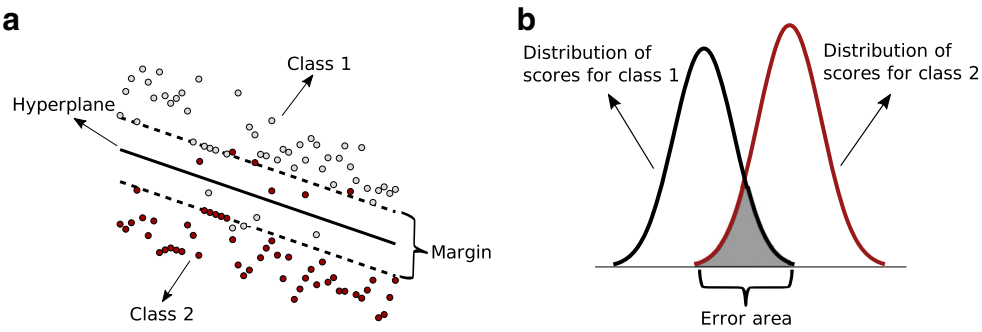


Fig. 8 **a** Representation of the SVM and **b** the SVM score distribution

the PD and aHC groups are considered for training. This experiment is to evaluate the suitability of the features to discriminate between Parkinson's patients and age-matched healthy controls. The third SVM is trained considering only the yHC speakers and PD patients. Both, age and PD are factors that affect the speech of elderly people, thus the difference between young speakers and PD patients should be larger than between young and healthy elderly people.

In order to analyze the results from the three experiments, the scores of the SVM are used to model the separability of the three classes, i.e, yHC vs. aHC, PD vs. aHC, and yHC vs. PD. These scores represent the distance of each data point to the separating hyperplane. Figure 8 shows a representation of the separation between two classes using an SVM and the probability density distribution of the distance of each data point to the separating hyperplane. The shadowed portion of the two distributions in Fig. 8b represents the probability of a sample to be misclassified. The "error area" is equivalent to the margin indicated in Fig. 8a.

Experiments and Results

Score Analysis

The SVM score analysis is performed for the three cases described above and training SVM with different feature vectors: (1) only phonation features, (2) only articulation features, and (3) the combination of both. Figure 9 shows the histograms and the fitted probability density distribution of the scores obtained from the phonation and articulation measures. The fitted distribution is based on a normal kernel function, and is evaluated at equally spaced points, that cover the range of the data. The SVM is trained considering features extracted from speakers of the the aHC and yHC groups. Both groups are statistically different when using phonation features ($t(98) = -6.81$, $p < 0.001$), articulation features ($t(98) = -11.11$, $p < 0.001$), and the combination of both sets ($t(98) = -11.98$, $p < 0.001$). The alpha level is set to 0.01 for all statistical tests. Figure 9 shows that there is a clearer separation between the fitted distributions when the articulation features are considered. These results confirm previous findings reported in the literature, where the deterioration in the articulatory capability in the speech of elderly people is described.

Figure 10 shows the fitted distributions for the PD patients and aHC speakers. These groups are not statistically different when training only with phonation features (($t(98) = -2.28$, $p = 0.025$)) and nor with the articulation features (($t(98) = -2.42$, $p = 0.017$)). However, the two groups are statistically different when the phonation and articulation features are combined (($t(98) = -3.50$, $p < 0.001$)). All the statistical tests were performed with an alpha value of 0.01. These results indicate that in order to model the speech impairments of people with PD, it is necessary to include information about their phonation and articulation capabilities, and it makes sense considering that PD affects among others, the vocal fold movement, the respiration process and the proper control of the articulator that are involved in the speech production process, e.g., tongue, lips, and jaw.

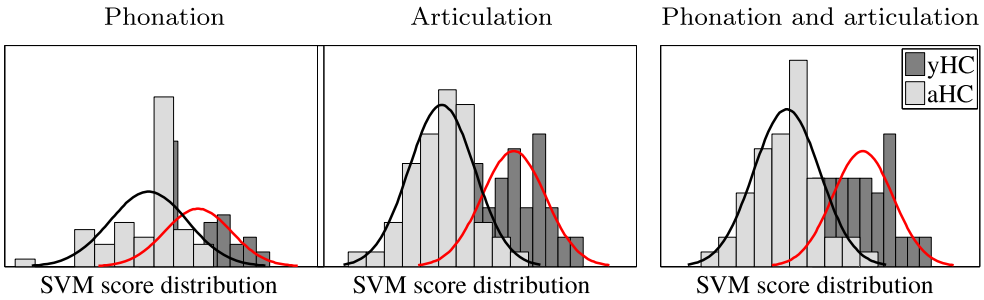


Fig. 9 Histograms and their corresponding fitted probability density distributions for the scores obtained from the yHC group (*dark-gray histograms with red curves*) and the aHC group (*light-gray histograms with black curves*)

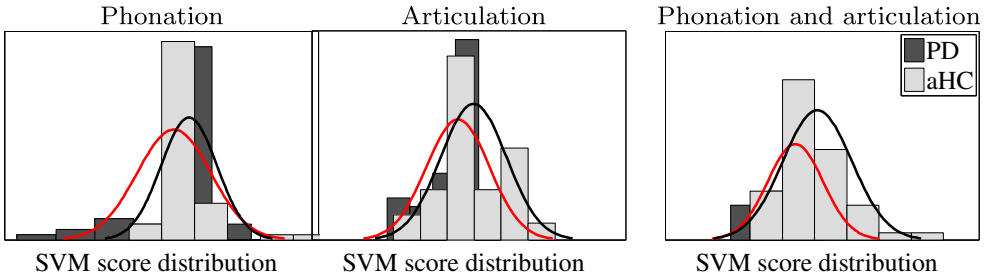


Fig. 10 Histograms and their corresponding fitted probability density distributions for the scores obtained from the PD group (*dark-gray histograms with red curves*) and the aHC group (*light-gray histograms with black curves*)

PD patients (PD) and young speakers (yHC) are statistically different when using phonation ($t(98) = -6.55$, $p < 0.001$) and articulation features ($t(98) = -13.84$, $p < 0.001$). With the combination of both feature sets the result shows also difference among groups ($t(98) = -6.36$, $p < 0.001$). The alpha level here is also set to 0.01 for the statistical tests. Figure 11 displays the fitted probability density distributions for the score vectors. Note that the articulatory measures are the most suitable to detect differences in speech of PD and yHC speakers, which confirms the results shown in Fig. 9.

Relevance Analysis

Feature selection consists of eliminating features with the highest linear correlation, i.e., features that provide the same or similar information. Phonation and articulation features are extracted from the utterances. 10-fold cross validation analysis is performed in order to compute a mean weighted vector $\bar{\boldsymbol{\rho}}$, with $\bar{\boldsymbol{\rho}}_k = 1/10 \sum_{i=1}^{10} \boldsymbol{\rho}_{ki}$, where $\boldsymbol{\rho}_{ki}$ is the weight of relevance of the k-th feature in the i-th fold. The original features are sorted according to $\bar{\boldsymbol{\rho}}$. Features with a correlation greater than 80% are eliminated considering the relevance order given by $\bar{\boldsymbol{\rho}}$.

In the case of phonation, a total of 480 features were extracted from the phonation of the five Spanish vowels and 309 were selected after the relevance analysis. The most relevant features were the shimmer, the APQ, the NNE and the PPQ. This result indicates that the stability of the vocal fold vibration is the most important characteristic in the phonation process at least to discriminate the three groups of speakers considered in this paper.

The set of measures computed for the articulation comprises 2289 features and a total of 1276 remained after the feature selection stage. The most relevant features were the first and second derivatives of the MFCCs. However, it is not clear whether a particular coefficient is more relevant than the others. This result indicates that at least to represent the articulatory capability of speakers based on sustained phonations, the MFCCs are the most suitable features. This result confirms previous findings reported in the literature where the capability of the MFCCs to model irregular movements in the vocal tract is shown [42]. Regarding the combination of the phonation and articulation features results on a 2769-dimensional feature vector, which is reduced to 1581 features after the relevance analysis. In this case, most of the features are from the articulation set (the first and second derivatives of the MFCCs) along with noise measures.

None of the Spanish vowels showed to be more relevant than the others. However, the mean and standard deviation computed for the features are always sorted at the top of the relevance vector. This behavior is depicted in Fig. 12 for the case of phonation, articulation, and the combination of both feature sets.

Binary SVM and Optimization of the Multi-class SVM

Three individual binary SVM were trained following a "one vs all strategy". Two scenarios are included: (1) the feature space is considered with the total number of features, and (2) the feature space is reduced by performing PCA-based feature selection following the relevance analysis introduced in "Relevance Analysis". Table 2 shows the accuracy, sensitivity, specificity, and the Area Under the ROC Curve (AUC)

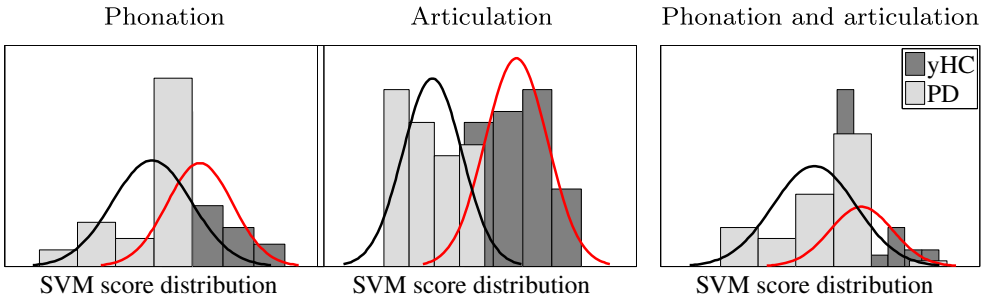


Fig. 11 Histograms and their corresponding fitted probability density distributions for the scores obtained from the yHC group (*dark-gray histograms with red curves*) and the PD group (*light-gray histograms with black curves*)

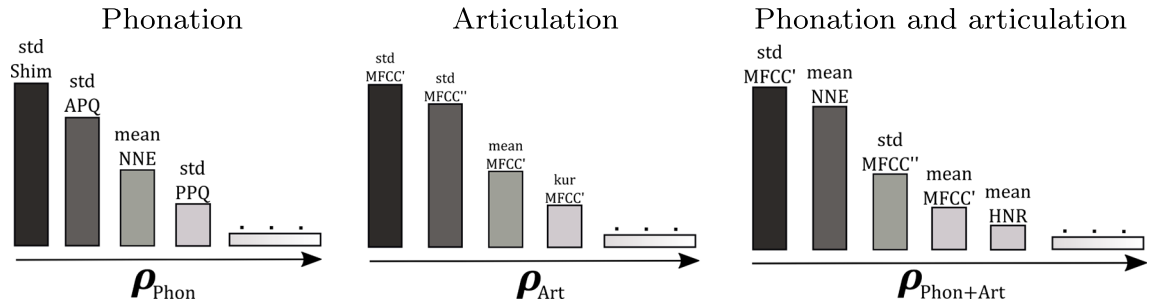


Fig. 12 Graphical representation of the relevance analysis for phonation, articulation, and the combination of both feature sets

obtained when training each SVM with the complete feature spaces (ROC stands for Receiver Operating Characteristic). The aim of this stage is to find the optimal meta-parameters (γ and C) for the subsequent multi-class SVM. It can be observed that highest accuracies are obtained discriminating the yHC group (81, 94, and 95%, for phonation, articulation, and the combination of both, respectively). Based on these results, it is expected the multi-class SVM to be able to discriminate young speakers from the others with a relatively high accuracy. When discriminating PD and aHC groups from the others, the accuracies are not high, indicating that aging is a confounding factor between healthy elderly speakers and people with Parkinson's disease. The best results obtained with no feature selection are compactly displayed in the ROC curves of Fig. 13.

The results obtained when PCA-based feature selection is considered (scenario 2) are displayed in Table 3. Note that these results are, in general, lower than those obtained when no feature selection is applied. These results could indicate that the correct discrimination of each class is a complex task and requires information from all of the features. With the aim of providing the reader more elements to analyze the two scenarios, with and without feature selection, both cases are also considered in the experiments with the multi-class SVM.

Figure 14 shows the obtained ROC curves for each binary SVM. In this case, the most relevant features are selected following the PCA-based approach. Note that in general, higher AUC values are obtained withe the articulation features. The only exception is the aHC group which exhibits higher AUC values when phonation and articulation features are merged.

After training the binary classifiers, the multi-class classification is performed considering the optimal meta-parameters γ and C. The aim of this experiment is to assess the influence of young speakers in the automatic analysis Parkinson's voice. Table 4A displays the confusion matrix obtained from the automatic classification of the three groups of speakers when the complete set of articulation features described in "Feature Extraction" is considered. Table 4B shows the performance of the multi-class SVM obtained when the PCA-based feature selection procedure is performed. The meta-parameters of the classifier, previously optimized in the binary-classification process, are also included in the table. Note that most of the misclassified PD patients are confused with speakers of the aHC group (Table 4A: 44%; Table 4B: 32%). Similarly, most of the errors discriminating aHC speakers are made with PD patients (Table 4A and B: 28%). In this case, the feature selection seems to be beneficial for the automatic discrimination

Table 2 Results (in %) for the binary SVM trained following a "one vs all" strategy

Features	Optimal parameters	SVM	ACC	SEN	SPE	AUC
Phonation	$C = 100, \gamma = 100$	PD vs All	70	56	76	0.70
		aHC vs All	70	55	78	0.71
		yHC vs All	81	68	90	0.91
Articulation	$C = 1, \gamma = 100$	PD vs All	77	65	84	0.84
		aHC vs All	71	55	83	0.76
		yHC vs All	94	90	96	0.96
Phonation and articulation	$C = 100, \gamma = 1000$	PD vs All	79	67	85	0.82
		aHC vs All	71	54	89	0.81
		yHC vs All	95	92	96	0.99

The complete feature space is considered, i.e., no PCA-based feature selection is performed. *ACC* accuracy, *SEN* sensibility, *SPE* specificity, *AUC* area under the ROC curve

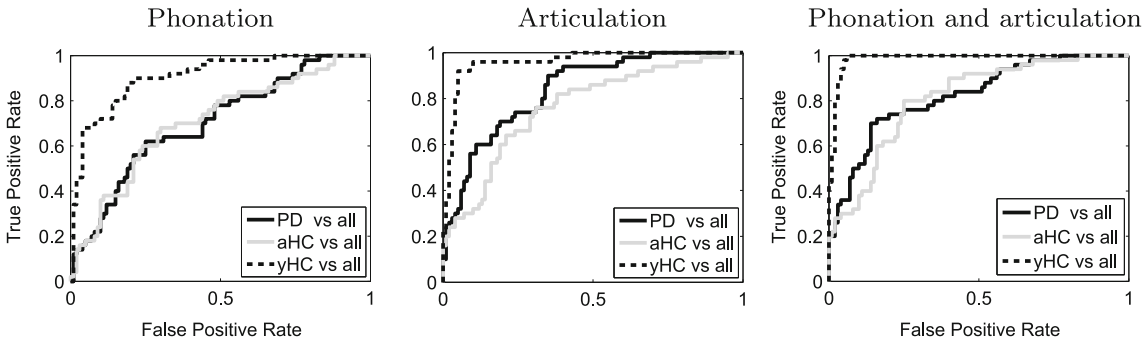


Fig. 13 ROC curves obtained with the "one vs all" strategy. No PCA-based feature selection is performed

of PD speakers because the accuracy improves from 50% (Table 4A) up to 66% (Table 4B). For the age-matched healthy speakers the performance of the classifier was lower with feature selection (60%) than with the complete feature set (66%). In both cases, the same number of aHC speakers are misclassified as PD patients (28%). The results show that feature selection improves the discrimination of patients and healthy speakers (aHC and yHC groups). Note also that the highest accuracy is obtained with young healthy speakers (yHC) in both scenarios (Table 4A and B). Further, the performance of the classifier is improved after feature selection (from 84 to 89%). The results obtained with the articulation features confirm previous observations made in "Score Analysis", where clear differences in the articulatory capability of young healthy speakers are observed with respect to elderly healthy people (aHC) and Parkinson's patients.

The results obtained with the phonation features are reported in Table 5. Both scenarios, with and without feature selection, are considered. Similarly to the results presented in Table 4, with the phonation features most of the PD patients are misclassified as elderly healthy controls (Table 5A: 52%; Table 5B: 22%) and most of the elderly speakers are misclassified as PD patients (Table 5A: 26%; Table 5B: 29%). In this case, the performance of the multi-class SVM improved for patients and aged healthy controls. Similarly to the articulation features, the discrimination between PD and aHC groups improved after performing the feature selection. The mis-classifications of aHC as yHC speakers also decreased. The classification of young speakers was the only case where the the accuracy decreased after feature selection, from 84 to 78%. Although this accuracy reduction, most of the mis-classifications were made with the aHC but not with PD speakers, which is positive if the aim is to have a low rate of false positives.

Besides the experiments with phonation and articulation features separately, both feature sets are merged into one representation space in order to evaluate the suitability of both speech dimensions to discriminate the three groups of speakers. The results of such a merging experiment are displayed in Table 6.

Note that there is an improvement in the discrimination of the three groups of speakers with respect to the results obtained with only phonation or articulation features.

Table 3 Results (in %) for the binary SVM trained following a "one vs all" strategy

Features	Optimal parameter	SVM	ACC	SEN	SPE	AUC
Phonation	$C = 100, \gamma = 100$	PD vs all	72	58	79	0.73
		aHC vs all	69	54	76	0.68
		yHC vs all	85	80	88	0.90
Articulation	$C = 0.1, \gamma = 1000$	PD vs all	79	69	83	0.83
		aHC vs all	69	52	81	0.79
		yHC vs all	93	95	92	0.96
Phonation	$C = 0.1, \gamma = 1000$	PD vs all	75	59	86	0.81
and		aHC vs all	71	55	87	0.76
articulation		yHC vs all	93	93	93	0.98

The reduced feature space is considered (PCA-based feature selection is performed). *ACC* accuracy, *SEN* sensibility, *SPE* specificity, *AUC* area under the ROC curve

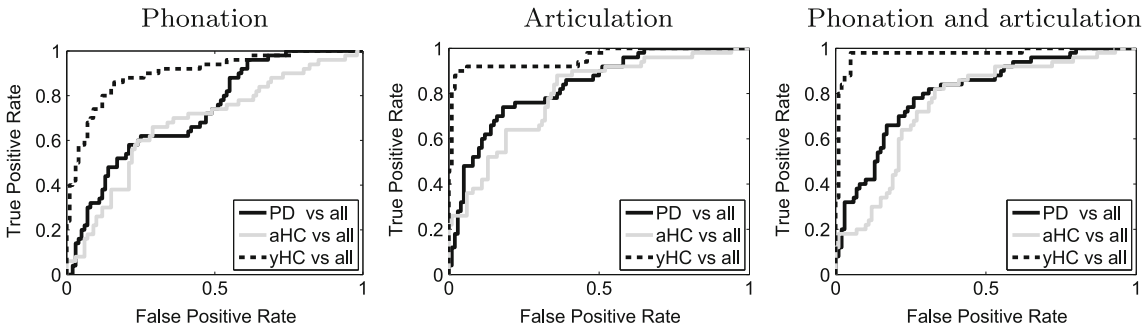


Fig. 14 ROC curves obtained with the "one vs all" strategy. PCA-based feature selection is performed

Regarding the impact of the feature selection process, in the case of the PD patients the accuracy improved from 54% (Table 6A) to 67% (Table 6B). For the aHC speakers, the performance was similar before and after the feature selection (Table 6A: 68%; Table 6B: 67%). The classification of the yHC speakers improved from 88 to 96% when feature selection is applied. These results indicate that feature selection is a good alternative to improve the automatic detection of Parkinson's patients when the control group includes age-matched and young speakers. As in the previous experiments, most of the misclassified speakers are PD patients and elderly healthy speakers. These results confirm, with experimental evidence, that age is a confounding factor for the automatic detection of Parkinson's disease.

Besides the aging influence analysis, the influence of the gender in the multi-class SVM is also studied. Table 7 shows the results obtained with the multi-class SVM trained with male and female speakers separately. In this case, phonation and articulation features are merged in one feature space. For both, male and female, most of the patients were misclassified in the aHC (56%). Table 7A shows the results obtained when only male are considered. It can be observed that most of the speakers in the aHC group are misclassified as patients (28%). Table 7B shows that when only female speakers are considered, there is an improvement in the detection of speakers of the aHC group (from 68 to 88%). The other results are similar compared to those obtained when female and male speaker are considered together. The results obtained in this experiment suggest that the accuracy of the system improves when only female speakers are considered. This behavior is not similar when only male speakers are considered. Further research with enough number of speakers per gender is required to find more conclusive results.

Cognitive-Inspired Classifier

Cognitive-inspired systems have been studied for decades [47, 48]. Recently, in [49] a special issue on brain-inspired

Table 4 Confusion matrix obtained with the articulation features. γ and C are optimized in the training stage performed with the binary SVMs

		Estimated class		
Optimal parameters	Target class	PD	aHC	yHC
A. Complete set of features				
$C = 1, \gamma = 1000$	PD	50	44	6
	aHC	28	66	6
	yHC	4	12	84
B. Feature selection				
$C = 0.1, \gamma = 1000$	PD	66	32	6
	aHC	28	60	12
	yHC	5	6	89

Results in %. C and γ are optimized on development

Table 5 Confusion matrix obtained with the phonation features. γ and C are optimized in the training stage performed with the binary SVMs

		Estimated class		
Optimal parameters	Target class	PD	aHC	yHC
A. Complete set of features				
$C = 100, \gamma = 100$	PD	40	52	8
	aHC	26	58	16
	yHC	2	14	84
B. Feature selection				
$C = 100, \gamma = 100$	PD	64	22	14
	aHC	29	63	8
	yHC	4	18	78

Results in %. C and γ are optimized on development

Table 6 Confusion matrix obtained with the merged phonation and articulation features. γ and C are optimized in the training stage performed with the binary SMVs

		Estimated class		
Optimal parameters	Target class	PD	aHC	yHC
A. Complete set of features				
$C = 100, \gamma = 1000$	PD	54	44	2
	aHC	26	68	6
	yHC	2	10	88
B. Feature selection				
$C = 0.1, \gamma = 1000$	PD	67	27	6
	aHC	27	67	6
	yHC	2	2	96

Results in %. C and γ are optimized on development

cognitive systems is presented. A total of 18 works are included in such an issue, which indicates the relevance of this topic in the state-of-the-art. The aim of such systems is to find mathematical representations of the way biological networks process information. One of the most widely studied system consists in neural networks (NN) which are to some extent designed to model the human brain. In this study, we limited the used of NN to a classification system based on a Multi-Layer Perceptron (MLP). A tri-class neural network (NN) is trained in order to compare it with respect to the best results obtained with the multi-class SVM. Previous works have shown the suitability of the multi-class NN for discrimination of emotional speech [50]. For these experiments, a neural network with three output

Table 7 Confusion matrix obtained merging the phonation and articulation features. Female and male speakers are considered separately

		Estimated class		
Optimal parameters	Target class	PD	aHC	yHC
A. Multi-class SVM trained with male				
$C = 0.1, \gamma = 1000$	PD	40	56	4
	aHC	28	68	4
	yHC	0	8	92
B. Multi-class SVM trained with female				
$C = 0.1, \gamma = 1000$	PD	40	56	4
	aHC	4	88	8
	yHC	4	12	84

Results in %. C and γ are optimized on development

units is used (PD patients, aHC, and yHC). The number of units of the hidden layer l is optimized trough a grid-search such that $l \in \{4, 10, 15, \ldots, 30\}$. The training process consists in determining the weight matrix $\boldsymbol{w}$ such that minimizes the error function $E(\boldsymbol{w})$ known as the cross-entropy loss function. For a standard multi-class classification, the error function is defined by the Eq. 19

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \ln\left[y_k(\boldsymbol{x}_n, \boldsymbol{w})\right], \tag{19}$$

where N is the number of inputs, K the number of classes, t_{kn} are the target values, $\boldsymbol{x}_n$ are the feature vectors, and $y_k(\boldsymbol{x}_n, \boldsymbol{w})$ is the output activation function used to compute the outputs y_k. In order to find the matrix $\boldsymbol{w}$ such that $E(\boldsymbol{w})$ is minimized, the gradient of the error function is found by means of the back propagation algorithm. During the optimization of the error function, a weight value has to be updated in the direction of the negative gradient of the error function. This procedure is illustrated in Eq. 20

$$w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E(w^{(\tau)}), \tag{20}$$

where τ indicates the iteration step, and η is the learning rate parameter such that $\eta > 0$. After updating $\boldsymbol{w}$, the gradient is computed again for the new weight and the process is repeated. After each step, the weight matrix is "moved" towards the greatest decreasing rate of the error function. The gradient is evaluated following the back propagation algorithm, which trains the NN for a given set of inputs $\boldsymbol{x}_n$ with a known classification targets t_k. The output of the NN is compared to the target values t_k and the error is computed. The weights of the NN are updated considering the computed error [45]. Figure 15 shows a diagram that summarizes the back propagation procedure. $\boldsymbol{x}$ is the feature vector which is the input to the first layer of the network. Each element of $\boldsymbol{x}$ represents an acoustic feature for each speaker in the database. The vector is forward propagated through the network (solid lines). At the end of the process, $\delta_k = y_k - t_k$ is calculated for all the output units and back propagated in the network (dashed lines). Afterwards, the weights of each input node are updated and the process is repeated until finding the minimum value of the error function.

Table 8 shows the performance of the tri-class NN when the complete set of phonation and articulation features are merged. Note that the highest performance was obtained for the young healthy group (98%). Conversely, for both PD patients and age-matched healthy controls most of the misclassified speakers are in the young healthy group. These

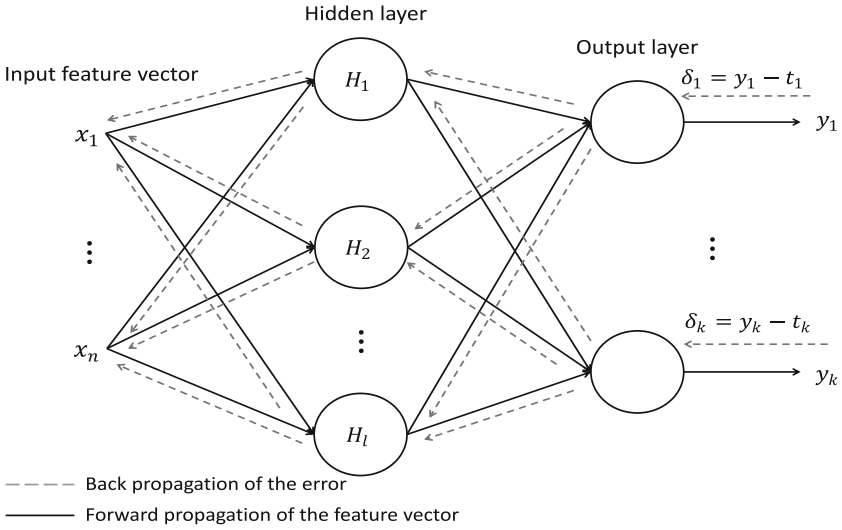


Fig. 15 Neural network with back propagation and k output classes

results indicate that the NN is more sensitive to the yHC class than to the other group of speakers. After the feature selection procedure, the performance of the classifier improved (Table 8A). For the PD patients the improvement is from 38 to 68%. The amount of PD patients misclassified as young speakers decreased from 50% (Table 8A) to 14% (Table 8B). In the case of the elderly healthy speakers, the accuracy increased from 32 to 52%. Although the performance for the aHC group is lower than in the multi-class SVM (Table 6B), most of the misclassified aHC speakers are confused with PD patients (Table 8B: 30%). As in the case of the multi-class SVM, the highest accuracy was obtained discriminating yHC speakers (92%). In general, the multi-class SVM exhibited better results than the tri-class NN in both scenarios: with and without feature selection. This can be explained considering that the multi-class SVM is more robust than the NN and its meta-parameters were optimized in a previous step based on a binary SVM.

Table 8 Confusion matrix obtained merging the phonation and articulation features and using a tri-class NN

		Estimated class		
Best	Target class	PD	aHC	yHC
A. Complete set of features				
$l = 25$	PD	38	12	50
	aHC	12	32	56
	yHC	2	2	98
B. Feature selection				
$l = 20$	PD	68	18	14
	aHC	30	52	18
	yHC	2	6	92

Conclusions

Sustained phonations of the five Spanish vowels uttered by three different groups of speakers: Parkinson's patients (PD), age-matched healthy controls (aHC), and young healthy speakers (yHC) are considered. The influence of PD in the phonation and articulation capabilities of the speakers is analyzed. Aging as a confounding factor to detect PD is analyzed considering the other two sets of speakers: 50 young healthy participants and 50 elderly healthy controls (with ages matched with respect to the PD group). Phonation and articulation measures are extracted from the voice signals in order to evaluate which of those speech dimensions (phonation and articulation) are more suitable to discriminate among the three groups of speakers. Several statistical tests are performed to evaluate whether there is significant difference between groups (PD vs. aHC, PD vs. yHC, and aHC vs. yHC). According to the results, phonatory and articulatory properties of the aHC and yHC groups are statistically different, thus the aging factor can be modeled considering each feature set separately or their combination. Similarly, when comparing PD with respect to yHC speakers, both speech dimensions are statistically different. However, when comparing PD vs. aHC speakers, each dimension is not statistically different. It is necessary to combine them in order to obtain statistical differences between those two groups. These results indicate that phonation and articulation capabilities of the speakers are impaired not only due to the presence of PD but also due to the aging process, thus in order to differentiate between PD and age-matched healthy control people, it is necessary to include more measurements and speech tasks like prosody and intelligibility extracted from read texts and monologues.

Feature selection with relevance analysis is performed. The resulting phonation and articulation measures are used

to model the speech of the speakers and the automatic discrimination among them is performed using a multi-class SVM with Gaussian kernel. The data are distributed into three groups: train, development, and test. The parameters of the classifiers are optimized on development to avoid over-fitted results. In all of the experiments (with phonation, articulation, and their combination), PD and aHC speakers are not separable while the detection of yHC speakers exhibited the highest accuracies in all of the cases. These results confirm those obtained with the statistical tests. Additionally, the results obtained when the phonation and articulation measures are merged were compared with respect to a tri-class neural network. The performance of the multi-class SVM was better than the NN; however, when feature selection is performed, similar results were achieved with both classifiers. These results indicate that it is possible to improve the detection of the pathology from speech when the feature selection stage is included in the automatic classification system.

To the best of our knowledge, this is the first paper introducing experimental evidence to support the fact that age matching is necessary to perform more accurate and robust evaluations of pathological speech signals. Additionally, the comparison among groups of speakers at different ages is necessary in order to understand the natural change in speech due to the aging process.

According to the findings reported in this paper, phonation and articulation features extracted from sustained vowels are only suitable to design a system to automatically discriminate between PD people and age-matched healthy controls. When the control group includes young speakers, it is necessary to consider other approaches. According to our preliminary experiments, the inclusion of features extracted from continuous speech, e.g., prosody, intelligibility, and articulation, could be enough to obtain satisfactory results.

We are currently working on a system to automatically discriminate among several kinds of diseases that affect different parts of the vocal tract (neurological: Parkinson's, organic: laryngeal cancer, and functional: cleft lip and palate) considering continuous speech recordings. Our main goal is to be able to objectively describe which measures are the most suitable to model each kind of disease.

Acknowledgments This research was partially funded by CODI at Universidad de Antioquia through the projects PRV16-2-01 and 2015-7683, and by COLCIENCIAS project no. 111556933858.

Compliance with Ethical Standards This study was partially funded by CODI at Universidad de Antioquia (grants number PRV16-2-01 and 2015-7683) and by COLCIENCIAS (grant number 111556933858).

Conflict of Interest The authors declare that they have no conflict of interest.

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

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Sataloff RT, Rosen DC, Hawkshaw M, Spiegel JR. The aging adult voice. J Voice. 1997;11(2):156–60.
2. de Rijk M. Prevalence of Parkinson's disease in Europe: a collaborative study of population-based cohorts. Neurology. 2000; 54:21–3.
3. Logemann JA, Fisher HB, Boshes B, Blonsky ER. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. J Speech Hear Disord. 1978;43(1):47–57.
4. Israel H. Age factor and the pattern of change in craniofacial structures. American J Anthropology. 1973;39(1):111–28.
5. Zaino C, Benventano T. Functional involutional and degenerative disorders. In: Zaino C and Benvetano T, editors. Radiographic examination of the oropharynx and esophagus. New York: Springer-Verlag; 1977.
6. Adams D. Age changes in oral structures. Dent Update. 1991;18(1):14–7.
7. Kahane J. Anatomic and physiologic changes in the aging peripheral speech mechanism. In: Beasley D and Davis G, editors. Aging communication process and disorders. New York: Grune & Stratton; 1981.
8. Benjamin BJ. Frequency variability in the aged voice. J Gerontol. 1981;36(6):722–6.
9. Steve AX, Deliyski D. Effects of aging on selected acoustic voice parameters: preliminary normative data and educational implications. Educ Gerontol. 2001;27(2):159–68.
10. Linville SE, Rens J. Vocal tract resonance analysis of aging voice using long-term average spectra. J Voice. 2001;15(3):323–30.
11. Ho AK, Iansek R, Marigliani C, Bradshaw JL, Gates S. Speech impairment in a large sample of patients with Parkinson's disease. Behav Neurol. 1999;11(3):131–7.
12. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. J Speech Lang Hear Res. 1969;12(2):246–69.
13. Hanson DG, Gerratt BR, Ward PH. Cinegraphic observations of laryngeal function in Parkinson's disease. Laryngoscope. 1984;94(3):348–53.
14. Orozco-Arroyave JR, Belalcázar-Bolaños EA, Arias-Londoño JD, Vargas-Bonilla JF, Skodda S, Rusz J, Daqrouq K, Hönig F, Nöth E. Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. IEEE J Biomedical Health Informatics. 2015;19(6):1820–28.
15. Ackermann H, Ziegler W. Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. J Neurol Neurosurg Psychiatry. 1991;54(12):1093–98.
16. Skodda S, Visser W, Schlegel U. Vowel articulation in Parkinson's disease. J Voice. 2011;25(4):467–72.
17. Orozco-Arroyave JR, Hönig F, Arias-Londoño JD, Vargas-Bonilla JF, Skodda S, Rusz J, Nöth E. Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's

disease, In: Proceedings of the 16th annual conference of the international speech communication association (INTERSPEECH). 2015, pp. 95–99.
18. Sapir S, Ramig LO, Spielman JL, Fox C. Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech. J Speech Lang Hear Res. 2010;53(1):114–25.
19. Rusz J, Cmejla R, Tykalova T, Ruzickova H, Klempir J, Majerova V, Picmausova J, Roth J, Ruzicka E. Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. J Acoust Soc Am. 2013;134 (3):2171–81.
20. Tsanas A, Little M, McSharry PE, Spielman J, Ramig LO, et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng. 2012;59(5):1264–71.
21. Tsanas A. Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning. United Kingdom: Oxford University; 2012.
22. Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng. 2009;56(4):1015–22.
23. Trail M, Fox C, Ramig LO, Sapir S, Howard J, Lai EC. Speech treatment for Parkinson's disease. NeuroRehabilitation. 2005;20(3):205–21.
24. Orozco-Arroyave JR, Belalcázar-Bolaños EA, Arias-Londoño JD, Vargas-Bonilla JF, Haderlein T, Nöth E. Phonation and articulation analysis of Spanish vowels for automatic detection of Parkinson's disease, In: Text, speech and dialogue, Springer; 2014, pp. 374–81.
25. Gómez-Vilda P, Rodellar-Biarge V, et al. Characterizing neurolgical disease from voice quality biomechanical analysis. Cogn Comput. 2013;5(4):399–425.
26. Deliyski D, Gress C. Intersystem reliability of MDVP for Windows 95/98 and DOS, In: Proceedings of the annual convention of the American speech-language-hearing association. San Antonio; 1998.
27. Goy H, Fernandes DN, Pichora-Fuller MK, Van Lieshout P. Normative voice data for younger and older adults. J Voice. 2013;27(5):545–55.
28. Torre P, Barlow JA. Age-related changes in acoustic characteristics of adult speech. J Commun Disord. 2009;42(5):324–33.
29. Boersma P, Weenink D. Praat, a system for doing phonetics by computer. Glot International. 2001;5(9/10):341–45.
30. Benjamin BJ. Phonological performance in gerontological speech. J Psycholinguist Res. 1982;1(11):159–67.
31. Pernambuco L, Espelt A, de Lima KC. Screening for voice disorders in older adults (RAVI)—part III: cutoff score and clinical consistency. J Voice. 2017;31(1):117.e17–117.e22.
32. Ben-Messaoud MA, Bouzid A, Ellouz N. A new biologically inspired fuzzy expert system-based voiced/unvoiced decision algorithm for speech enhancement. Cogn Comput. 2016;8(3):478–93.
33. Siegert I, Philippou-Hübner D, Hartmann K, Böck R, Wendemuth A. Investigation of speaker group-dependent modelling for recognition of affective states from speech. Cogn Comput. 2014;6(4):892–913.
34. Travieso CM, Alonso JB. Special issue on advanced cognitive systems based on nonlinear analysis. Cogn Comput. 2013;5(4):397–8.
35. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stern MB, Dodel R, et al. Movement disorder society-sponsored revision of the unified Parkinsons disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord. 2008;23 (15):2129–70.
36. Benesty J, Mohan S, Yiteng HE. Springer Handhook of Speech processing. Springer-Verlag; 2008.
37. Kasuya H, Ebihara S, Chiba T, Konno T. Characteristics of pitch period and amplitude perturbations in speech of patients with laryngeal cancer. Electron Commun Jpn (Part I: Communications). 1982;65(5):11–9.
38. Yumoto E, Gould WJ, Baer T. Harmonics-to-noise ratio as an index of the degree of hoarseness. J Acoust Soc Am. 1982;71(6):1544–50.
39. de Krom G. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. J Speech Lang Hear Res. 1993;36(2):254–66.
40. Kasuya H, Ogawa S, Mashima K, Ebihara S. Normalized noise energy as an acoustic measure to evaluate pathologic voice. J Acoust Soc Am. 1986;80(5):1329–34.
41. Michaelis D, Gramss T, Strube HW. Glottal-to-noise excitation ratio–a new measure for describing pathological voices. Acta Acustica United with Acustica. 1997;83(4):700–6.
42. Godino-Llorente JI, Gomez-Vilda P, Blanco-Velasco M. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. IEEE Trans Biomed Eng. 2006;53(10):1943–53.
43. Orozco-Arroyave JR, Murillo-Rendón S, Álvarez-Meza AM, Arias-Londoño JD, Delgado-Trejos E, Bonilla-Vargas JF, Castellanos-Domínguez CG. Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection. In: Proceedings of the 11th annual conference of the international speech communication association (INTERSPEECH). 2011, pp. 529–532.
44. Daza-Santacoloma G, Arias-Londoño JD, Godino-Llorente JI, Sáenz-Lechón N, Osma-Ruíz V, Castellanos-Dominguez CG. Dynamic feature extraction: an application to voice pathology detection. Intelligent Automation & Soft Computing. 2009;15(4):667–82.
45. Bishop CM. Pattern Recognition and Machine Learning, 1st edn ser. Information Science and Statistics. Springer-Verlag; 2007.
46. Orozco-Arroyave JR. Analysis of speech of people with Parkinsons disease. Germany: Logos Verlag Berlin; 2016.
47. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5(4):115–33.
48. Rosenblatt F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. DTIC Document, Tech. Rep.; 1961.
49. Luo B, Hussain A, Mahmud M, Tang J. Advances in brain-inspired cognitive systems. Cognitive Computation. 2016;8(5): 795–6.
50. Henríquez P, Alonso JB, Ferrer MA, Travieso CM, Orozco-Arroyave JR. Global selection of features for nonlinear dynamics characterization of emotional speech. Cognitive Computation. 2013;5(4):517–25.