Speech impairment in Parkinson’s disease: acoustic analysis of unvoiced consonants in Italian native speakers

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 25, 2021, accepted December 6, 2021, date of publication December 14, 2021,
date of current version December 24, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3135626*

# Speech Impairment in Parkinson's Disease: Acoustic Analysis of Unvoiced Consonants in Italian Native Speakers

**FEDERICA AMATO[1], LUIGI BORZÌ[1], (Member, IEEE), GABRIELLA OLMO[1], (Senior Member, IEEE), CARLO ALBERTO ARTUSI[2], GABRIELE IMBALZANO[2], AND LEONARDO LOPIANO[2]**

[1]Department of Computer Engineering, Politecnico di Torino, 10129 Torino, Italy
[2]Division of Neurology, A.O.U. Città della Salute e della Scienza di Torino, 10126 Turin, Italy

Corresponding author: Federica Amato (federica.amato@polito.it)

**ABSTRACT** The study of the influence of Parkinson's Disease (PD) on vocal signals has received much attention over the last decades. Increasing interest has been devoted to articulation and acoustic characterization of different phonemes. **Method**: In this study we propose the analysis of the Transition Regions (TR) of specific phonetic groups to model the loss of motor control and the difficulty to start/stop movements, typical of PD patients. For this purpose, we extracted 60 features from pre-processed vocal signals and used them as input to several machine learning models. We employed two data sets, containing samples from Italian native speakers, for training and testing. The first dataset - 28 PD patients and 22 Healthy Control (HC) - included recordings in optimal conditions, while in the second one - 26 PD patients and 18 HC - signals were collected at home, using non-professional microphones. **Results**: We optimized two support vector machine models for the application in controlled noise conditions and home environments, achieving $98\% \pm 1.1$ and $88\% \pm 2.8$ accuracy in 10-fold cross-validation, respectively. **Conclusion**: This study confirms the high capability of the TRs to discriminate between PD patients and healthy controls, and the feasibility of automatic PD assessment using voice recordings. Moreover, the promising performance of the implemented model discloses the option of voice processing using low-cost devices and domestic recordings, possibly self-managed by the patients themselves.

**INDEX TERMS** Italian native speakers, Parkinson's disease, support vector machine, tele-health, unvoiced consonants, voice analysis, classification, machine learning.

## I. INTRODUCTION

Parkinson's Disease (PD) is a chronic and progressive disorder affecting 1% of the over 60 population worldwide, and it is expected to interests more than 9 million people in industrialized nations by 2030 [1]. This neurodegenerative condition alters the functions of the basal ganglia, and leads to a progressive loss of dopaminergic neurons, especially in the *substantia nigra* of the midbrain [2]. Patients with Parkinson's Disease (PDP)s manifest a broad spectrum of clinical symptoms, including bradykinesia, rigidity, tremor at rest, postural instability, sleep disorders, and speech impediment [3], [4]. This latter is receiving an increasing attention in the scientific community, due to the enormous amount of clinical information embedded in the vocal signal, despite the simple data collection modality. Indeed, the speech production is accomplished through synergistic articulating movements that shape the excitation source to convey the final sound [5]. The excitation source is the fundamental element of vocal production, and can be voiced, unvoiced, or a mixture of both [5]. In the first case, the sound is produced by forcing air through the vocal folds, which vibrate and generate a quasi-periodic signal. In the second case, there is no constriction of the vocal folds, and the airflow arrives unaltered to the articulating elements, where the final sound is created by forcing air through teeth, lips, and tongue [6]. During the speech production, the speaker merges these sounds to form phonemes, words, and phrases through a continuous alternation of voiced and unvoiced traits. This complex process, besides achieving the main objective of transmitting

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Kamrul Hasan.

information, incorporates a large amount of data of clinical interest. These data can be extracted and input to Machine Learning (ML) algorithms, implementing a useful tool to support the clinical practice.

Although speech analysis finds applications in any pathology that directly or indirectly affects the vocal apparatus, it is particularly effective in PD, as almost 90% of the affected population manifest alterations in speech production [7], [8]. Moreover, PD is known to have a prodromal phase during which neuro-degeneration is already underway, but cardinal symptoms have not manifested yet [9]. Speech impairment is known to occur up to 10 years earlier than cardinal manifestations, thus it can contribute to the early diagnosis of the disease [9], [10]. Speech impediments in PDPs are usually gathered under the general term of hypokinetic dysarthria and affect in different ways the three dimensions of speech (i.e. phonation, articulation, and prosody) [11]. Most of the research community tends to focus on phonation to assess the patient's ability to force air from the lungs to the vocal folds and make them vibrate to produce sounds [2]. However, it is well known that sustained vowels are an over-simplistic task, which does not include fluctuations in vocal characteristics such as voice onset, terminations, and breaks [12]. This over-simplification is directly linked to a reduced discriminatory capability: according to the comparison between phonatory and articulation approaches described in [12], the use of articulation features together with ML techniques maximizes the performance of PD automatic detection models, with accuracy ranging from 80% to 95%.

Besides phonation and articulation, impairments in the prosody of PDPs have been observed as well. Prosody studies mainly focus on speech rate, pause, intonation, and general communication skills of people [2]. As a result, the vocal signal analysis is performed at high level and the extracted parameters can be influenced by the data collection modality. Indeed, it is well known that anxiety and state of alert influence PD symptoms [13]. Hence, recordings performed in a controlled environment may not yield a realistic representation of the actual vocal alterations of the patient, as experienced in daily life. On the other hand, a detailed articulation analysis can investigate more specific aspects, less prone to variations due to the patient's emotional state. In more detail, features can be extracted from different types of sound regions and can be related to the speed or acceleration of articulation elements [12]. TRs between segments can be employed to describe the patient's ability to initiate and stop movements; the impaired articulation of different phonemes can be measured to investigate how the disease affects the regions of the phonation apparatus.

However, the aspects that contribute to the strong discriminatory potential of articulation analysis also lead to high complexity. While the analysis of consonants contains more information than vowels, the pronunciation of such phonemes varies depending on the language being considered [12], [14]. In addition, a detailed analysis of phonemes requires the use of more precise measurements than a prosody study. Moreover, although recent works [12], [15] investigated the importance of the distinct phonemic groups for the automatic identification of PD, a specific set of phonemes that is language-independent and has a proven high correlation with the disease is not available yet.

The physiological motivation behind the TRs analysis is the effect of the lack of coordination, typical of PDPs, in the use of the source glottal [16], [17]. Indeed, the direct visualization of the vocal fold vibration by video laryngoscopy [11] revealed incomplete glottis closure due to impaired vocal fold abduction and bowing. Additionally, asymmetry in vocal fold closure and arytenoid cartilage position and movement have also been described. The motor impairment deriving from this alteration can manifest in various manners. An example is the phenomenon of *voicing leakage* [17]: after the production of a voiced sound, PDPs face difficulties in interrupting the vocal fold movement; thereafter a partial vibration is perceived in lieu of the regular phonation interruption. Another consequence is the *spirantization*, a speech impediment that occurs when, due to incomplete closure of the vocal folds, air escapes during what should be a silent interval [18], implying a perceivable distortion of unvoiced consonants. For example, a /t/ spirantized by PDPs may sound more like /s/ [18]. Thereafter, in this work we aim to assess the effectiveness of an acoustic analysis based on the study of TRs between unvoiced consonants and the adjacent voiced segments, to investigate which phonemes of the Italian language are mostly affected by hypokinetic dysarthria, and which features are most suitable for characterizing them employing both optimal and sub-optimal recordings. From an engineering perspective, we believe that the detailed analysis of these alterations can help differentiating PDPs and HCs. Moreover, the investigation of the phonetic mis-articulation can provide enormous support to speech therapists during the development of a rehabilitation therapy tailored for a single patient, as well as during the follow-up stage.

The remainder of this paper is organized as follows. In section II we review automatic methodologies for PDPs speech analysis. In section III we address the employed data sets as well as the feature extraction and selection methods, and the classification model. In section IV we describe the classification performance and the analysis of our findings. Finally, in section V we draw conclusions and propose future improvements for the present model.

## II. RELATED WORK

The automatic identification of PD through the analysis of vocal recordings has gained increasing attention over the last decades. Interest has been recently devoted to the articulation approach and the acoustic characterization of different phonemes. In this context, [19] investigated the properties of fricatives produced by PDPs, as these consonants are commonly mispronounced in patients with dysarthria [20]. The authors analyzed a corpus including 10 PDPs and 9 HCs repeating two English words (*sigh and shy*) ten times.

The acoustic measures included duration, intensity, and four spectral moments. In fact, PDPs' speech is characterized by reduced segments length, and this is particularly evident in consonants. Intensity measures the difference between the fricative and the following segment, hence the ability of the patient to perform a complex sequencing of movements. Finally, spectral moments evaluate the co-articulation, hence the coordination between successive gestures. Despite the absence of a classification step, statistical analysis denoted the high potential of the features extracted. However, due to the reduced size of the dataset, the results can only be considered preliminary.

The relevance of nasal consonants in the automatic identification of PDPs was investigated in [21]. In this work, the authors explored the reliability of features extracted from the sustained voiced consonant /m/ of 40 Australian native English speakers (18 PDPs and 22 HCs). The parameters were fed into a Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) to differentiate PDPs and HCs, achieving 93% classification accuracy in the Leave One Out (LOO) Cross Validation (CV). Moreover, the Spearman correlation analysis showed that the features extracted were highly correlated to the Movement Disorder Society revised version of the Unified Parkinson's disease rating scale motor score (MDS-UPDRS-III). However, the performance of the algorithm is referred only to the LOO-CV, with no mention of accuracy achievable on new, previously unseen samples.

The possibility of employing specific phonemes in the early identification of PD has been investigated in [22], whose authors developed a diadochokinesis-based system considering articulation features of occlusive consonants. They extracted temporal and spectral parameters from the Voice Onset Time (VOT) segments of the /ka/, /ta/, /pa/ syllables, using a dataset composed of 27 Spanish PDPs and 27 age-matched Spanish HCs. The occlusive consonant /k/ exhibited the highest discrimination capability, reaching a classification accuracy of 94.4% in the case of LOO-CV and 92.2% using 10-fold CV. Also in this work, no experiment has been performed on an independent test set.

More recently, [17] analyzed the importance of distinct phonemic groups to discriminate PDPs from HCs. Starting from the assumption that different acoustic segments have different relevance, the authors proposed a method based on Perceptual Linear Prediction (PLP) features and Gaussian Mixture Models (GMM)-Universal Background Models (UBM) classifiers to investigate the importance of different phonemes. To this end, 5 corpora including Spanish and Czech native speakers were employed. The cross validation results reached an accuracy ranging from 85% to 94%(11-fold CV), while cross-corpora trials yielded an accuracy between 75% and 82%. The post-hoc analysis of the results suggested that occlusives, vowels, and fricatives are the most relevant acoustic segments in the considered languages. However, as also stated by the authors themselves, the Czech dataset is made up only of male speakers and PDPs are in an early stage of the disease, which prevents from considering the cross-language results exhaustive. In a subsequent work [1], the authors introduced for the first time the analysis of transition between specific phonemes and evaluated their influence in the detection of PD. In this experiment, they developed a model employing GMM-UBM classifiers and PLP as features extracted from relevant articulation moments, such as bursts, transitions between vowel and consonants, or the beginning and end of the glottal activity. The achieved accuracy in the Czech dataset was $94\pm1\%$, while the best cross-corpora accuracy was $82 \pm 13\%$. In both cases, the speech task was the Diadochokinetic (DDK) task. Also, the analysis demonstrated the influence of the language on the models performance.

In this context, the first objective of this work is to assess the effectiveness of an acoustic analysis based on the study of TRs between unvoiced consonants and the adjacent voiced segments. The second objective is to investigate which phonemes of the Italian language are mostly affected by hypokinetic dysarthria and which features are the most suitable for characterizing them. This task will convey valuable information about phonetic groups with the highest discriminating capability. This could enhance the identification of a reduced phonetically balanced speech task that can minimize the effort required to the patients, and the extension of the set of features used for the description of TR proposed in [1].

Finally, the study aims to assess whether it is possible to rely on recordings made independently by PDPs in their home environment without supervision, and whether it is possible to extract information from such recordings. We believe that the development of a system that requires a minimum effort to the patient, easy to use, and low cost can help in the patient's follow-up at home and can be employed by neurologists to monitor the progression of the disease. Such a system may support the clinical practice in patients' follow-up, while minimizing the bias in voice analysis due to the non-comfortable setting (e.g. hospital environment).

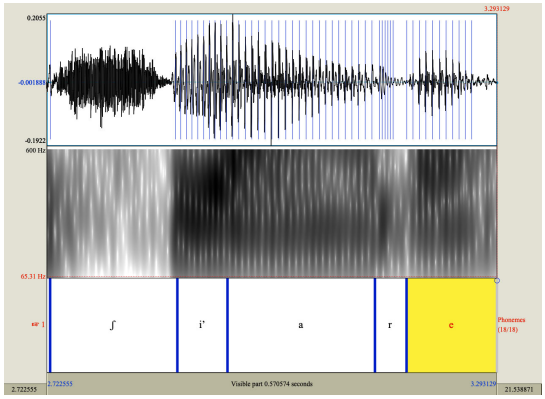
## III. MATERIALS AND METHODS

In this section, we describe the employed datasets as well as the algorithms implemented for PDPs' voice classification.

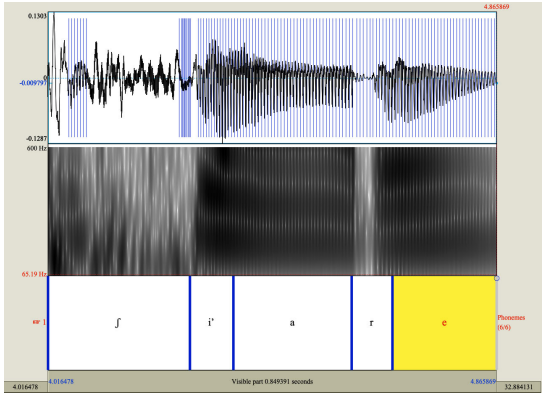
First of all, we want to identify the most suitable features to analyze TRs between unvoiced consonants and the adjacent voiced segments in the Italian language, and test whether different phonemes requires different features to be effectively characterized (section I). To this end, after performing a robust pre-processing, described in section III-B, we devoted much effort to feature extraction and selection, in order to identify a compact set of features able to quantify the phonetic mis-articulation. As an example of the abnormalities investigated in this study, in Fig. 1 we present the vocal signal, its spectrogram obtained employing the Praat default parameters, and the phonetic transcription of the Italian word *sciare* pronounced by a PDP and an age-matched HC. It can be appreciated that the spectrogram region related to the Italian unvoiced phoneme /ʃ/ is clearly altered in the PDP

**TABLE 1. List of phonemically balanced phrases employed in the present study. The original Italian sentence and the corresponding translation into the English language is reported.**

| Phrase ID | Phrase - Italian | Phrase - English translation |
|---|---|---|
| 0 | Oggi è una bella giornata per sciare | Today is a beautiful day for skiing. |
| 1 | Voglio una maglia di lana color ocra. | I want an ocher wool sweater. |
| 2 | Il motociclista attraversò una strada stretta di montagna. | The biker crossed a narrow mountain road. |
| 3 | Patrizia ha pranzato a casa di Fabio. | Patrizia had lunch at Fabio's house |
| 4 | Questo è il tuo cappello? | Is this your hat? |
| 5 | Dopo vieni a casa? | Will you come home later ? |
| 6 | La televisione funziona? | Does the television work? |
| 7 | Non posso aiutarti? | Can't I help you? |
| 8 | Marco non è partito | Marco did not leave. |
| 9 | Il medico non è impegnato. | The doctor is not busy. |




(a) *HC, gender:female, age:68*




(b) *PD, gender:female, age:67*

**FIGURE 1. Vocal signal, spectrogram and phonetic transcription of the Italian word *sciare* pronounced by a PD patient and an age-matched HC.**

case. Figure 2 depicts a simple flowchart to provide a general overview of the workflow.

### A. DATASET DESCRIPTION

#### 1) MAIN DATABASE

The main database employed for this study is a public corpus [23] made up of 65 Italian native speakers, including 15 young HCs (age $20.8 \pm 2.65$), 22 elderly HCs (age $67.09 \pm 5.16$) and 28 PDPs (age $67.21 \pm 8.73$). None of the HC reported speech or language disorders, and all PDPs received their usual anti-parkinsonian treatment. The Hoehn and Yahr (H&Y) was $< 4$, except for three patients (one classified as stage 5 and two as stage 4). All the recordings were performed with professional microphones in a quiet echo-free room. The participants were asked to execute a set of tasks, including reading of a phonetically balanced text, execution of the syllables /pa/ and /ta/, phonation of the vowels /a/, /e/, /i/, /o/, /u/, reading of a list of phonetically balanced words, and finally reading of a list of phonetically balanced sentences. For this specific application, we employed a subset of this dataset composed of the entire PD population and the 22 age-matched elderly controls. The task considered for the model development was the reading of the list of sentences, whose transcription and translation into the English language is reported in Table 1

#### 2) ADDITIONAL DATASET

Data from the first dataset was recorded in optimal conditions and this can be hardly reproduced in real-world scenarios. Thus, we collected a second database registered under sub-optimal conditions. We recorded data from 44 volunteers (26 PDPs, age $71.7 \pm 7.39$ and 18 HCs, age $65.5 \pm 8.42$) enrolled at *A.O.U. Città della Salute e della Scienza di Torino*, *Associazione Amici Parkinsoniani Piemonte Onlus* and *Imperia Hospital*. The inclusion criteria were: a clinical diagnosis of idiopathic PD with vocal signs and symptoms; no major cognitive impairment or other conditions preventing the patient from correctly accomplishing the task.

The collection of this database was performed through a web application that guided the users through the execution of the same tasks encompassed in the main dataset.

The data collection has been conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the *A.O.U Città della Salute e della Scienza di Torino* (approval number 00384/2020). Participants received detailed information on the study purposes and execution, and informed consent for observational study was obtained. Demographic and clinical data were noted anonymously.

### B. PRE-PROCESSING

This section describes the pre-processing steps carried out to ease the extraction of specific information from vocal signals. The entire analysis was performed through the software Praat and applied to each sentence listed in Table 1.

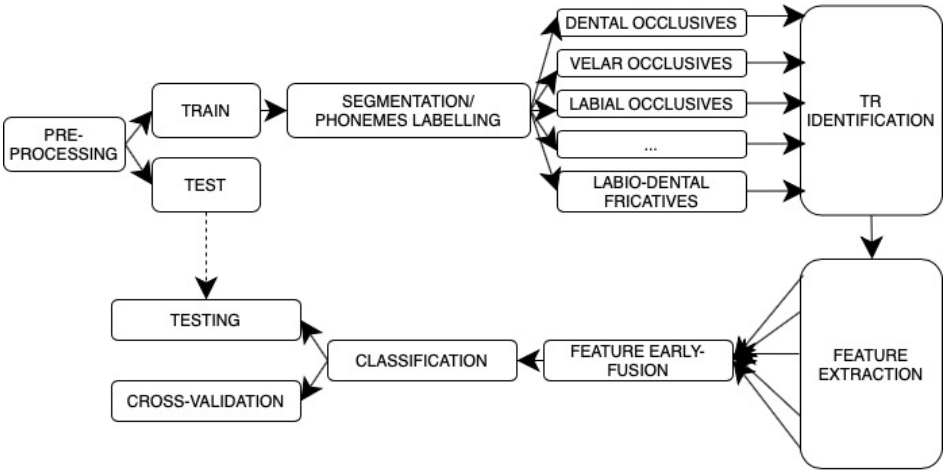



**FIGURE 2.** Work flow scheme.

**TABLE 2.** Overview of the set of features employed in the present work.

| Feature name | Information retrieved | Study |
|---|---|---|
| RASTA-PLP | Abnormalities in the articulation of specific sounds | [1], [17] |
| Spectral moments: mean | Inability to promptly interrupt/start vocal fold vibration | [19], [22] |
| Spectral moments: standard deviation | Inability to promptly interrupt/start vocal fold vibration | [19] , [22] |
| Spectral moments: skewness | Inability to promptly interrupt/start vocal fold vibration | [19] , [22] |
| Spectral moments: kurtosis | Inability to promptly interrupt/start vocal fold vibration | [19] , [22] |
| ETS | Inability to promptly interrupt/start vocal fold vibration | present study |
| MFCC | Subtle changes in the motion of articulators | [1] |
| DFA | Lack of coordination in vocal fold vibration | present study |
| Intensity difference | Reduced phonetic contrast | [19] |
| Duration ratio | Increased speech rate, reduced unvoiced consonant duration | [19] |

The recordings in the employed dataset were characterized by various sampling rates; hence, they were firstly down-sampled to 16 kHz to maintain similar spectral conditions. Thereafter, a de-noising filter with Praat default hyperparameters was applied to each signal, and their amplitude was normalized in the range [0, 1] to prevent the speaker-microphone distance from affecting the model. It is worth noticing that the visual and acoustic signal examination indicated the absence either of initial or final silence regions, hence no further preparatory steps were required. Then, we employed the Praat software to detect voiced regions' start and end-point, manually labelling each segment with the corresponding transcription. To perform the analysis of the TRs between unvoiced consonants and the adjacent voiced segments, we manually detected the regions corresponding to the transition between unvoiced consonants and a voiced segment. The use of an automatic segmentation system would have introduced a bias in the results. In fact, tools for automatic segmentation are characterized by an intrinsic error, which becomes even more evident in the case of PDPs, whose speech is affected by several alterations. Hence, we automatically identified 160 ms long windows centred on the edge of each chunk. According to [24], such window size allows to perform an in-depth analysis of the transient regions.

### C. FEATURE EXTRACTION

In accordance to the main objectives of this work (section I), the feature extraction procedure aims to identify a set of features able to embody the vocal alterations characterizing PDPs through the analysis of the transition from unvoiced consonants to the contiguous voiced region. More in detail, we extracted a set of parameters encompassing features previously involved in phonetic analysis as well as novel parameters which have the potential to describe the alteration in the TR. Table 2 presents an overview of the features used for the analysis. In the following, we provide a detailed description of the features employed. Relative Spectral - Perceptual Linear Prediction (RASTA-PLP) coefficients have been widely applied in the phonemic analysis due to their ability to provide information about acceleration and velocity of the articulators during speech production [1]. In this work, after the pre-processing steps, we divided each TR into 15ms with 50% overlap frame length [17]. Then we evaluated 13 RASTA-PLP coefficients for each frame together with their first and second derivative [25], grouped them into one feature vector, and calculated four statistics of these vectors (i.e. mean value, standard deviation, kurtosis, and skewness). As for the *spectral moments*, they have been employed in several studies for the characterization of dysarthric speech [19], [20]. We employed four spectral

moments: *mean spectral peak*, which measures whether most energy is concentrated in a small band or dispersed over a wider range; *spectral standard deviation*, which is a measure of deviation of the spectrum frequencies from the centre of gravity; *spectral skewness*, which measures the shape of the spectrum below the mean peak compared to the frequencies above it; *spectral kurtosis*, which describes the weakness of the energy distribution, with positive and negative values suggesting well defined spectral peaks and a flat distribution [20].

To further characterize the spectral differences in the TR, we introduced a novel parameter: the Energy Transition Slope (ETS). Based on the assumption that PDPs hardly perform rapid movements, we expect the energy contour in the unvoiced/voiced switch to exhibit a more flattened curve with respect to HCs. Hence, we evaluated the energy contour in the TR through a first-order polynomial and employed the slope of the obtained curve to embody this alteration.

Mel Frequency Cepstral Coefficients (MFCC)s mimic the efficient filtering capability of the human ear and have been widely applied to speaker identification and biomedical voice assessments [8], [26]. Although they do not admit a clear physical interpretation, they can detect subtle changes in the motion of the articulators (tongue, lips) [26] and can provide pivotal information on the impairment in the TR. As for the RASTA-PLP coefficients, after the pre-processing steps, we divided each TR into 15ms with 50% overlap length frame length; hence we evaluated 13 MFCCs for each frame together with their first and second derivative [22], [25], grouped them into one feature vector, and calculated four statistics for each array. Moreover, to further describe abnormalities in the unvoiced/voiced switch, we introduced the Detrended fluctuation analysis (DFA), which is usually employed to quantify the degree of stochastic self-similarity in the turbulent noise [8]. Indeed, due to the lack of control and coordination in the vocal fold, we expect the TRs to present turbulent disturbances and, consequently, an increased value of the DFA coefficient [27]. Finally, we employed the *intensity difference* and the *duration ratio* to measure the differences between the first region of the analyzed segment (i.e. the unvoiced consonant) and the subsequent voiced area. According to [19], the intensity difference between average values of the unvoiced consonant and the following vocal nucleus can reflect the reduced phonetic contrast, often described as *blurred speech*. As for duration ratio, it measures the ratio between the lengths of the unvoiced consonant and the adjacent voiced tract to quantify the reduced unvoiced consonant duration, which usually characterizes PDPs.

Since the gender of the speaker has a non-negligible influence on some speech characteristics, before performing feature selection, we combined vocal parameters with a covariate indicating whether the sample belongs to a male or a female subject, in order to pursue a feature selection procedure that takes into account gender-specific differences.

Given that different features exhibit different ranges, we applied the *z-score normalization* to the whole feature set, consisting in removing the mean value and dividing by the standard deviation. This, besides being a general good practice, is particularly important if Euclidean distances have to be computed in subsequent analysis (e.g. similarity measures).

### D. FEATURE SELECTION

Feature selection was employed to decrease the dimensionality of the input variables, reduce the computational cost of the model, and boost the performance of the prediction. We addressed the *Boruta* feature selection approach due to its successful applications in various domains [8]. The algorithm is a wrapper method based on a Random Forest (RF) classification algorithm, which aims to find all important variables by comparing the relevance of the real features to that of the random probes [28]. The chief assumption under Boruta's algorithm is that adding randomness to the system and analysing its impact on the model can highlight the most significant features [28]. For each input variable the algorithm creates a *shadow attribute*, which is obtained by shuffling the values of the original attribute across objects. Then, the RF classifier is trained with this extended set of features, the classification phase is performed, and the relevance of each attribute is computed. The importance of a shadow attribute can be non-zero only due to random fluctuations. Thus the relevance of shadow attributes is used as a reference to identify the smallest subset of relevant features [28].

### E. CLASSIFICATION

We implemented a model for the automatic binary classification between PDPs and HCs, devoting much attention to the importance of distinct phonetic groups and the influence of different recording conditions. In Experiment 1, we employed only samples belonging to the first database and we tested the model performance in optimal conditions, emulating the outpatient environment, when it is more likely to have professional equipment available. In Experiment 2, we introduced the additional database to test the suitability of voice data recorded in sub-optimal conditions, with low-cost equipment and in the absence of external supervision such as in the home environment.

In Experiment 1, in order to avoid weak generalization capability of the model, we randomly split the database into two subsets: 80% to be used during the training/validation phase and 20% to be used as the test set. The two sub-groups were chosen in such a way as to guarantee speaker independence (i.e., all sentences of the same speaker are either in training or in test set, but not split between the two). It is worth noting that we implemented feature selection, model selection and model optimization on the training/validation set, while the remaining 20% of subjects was employed only during the testing phase, without any further optimization.

A similar procedure was applied in Experiment 2. We randomly split both the main and the additional database into two

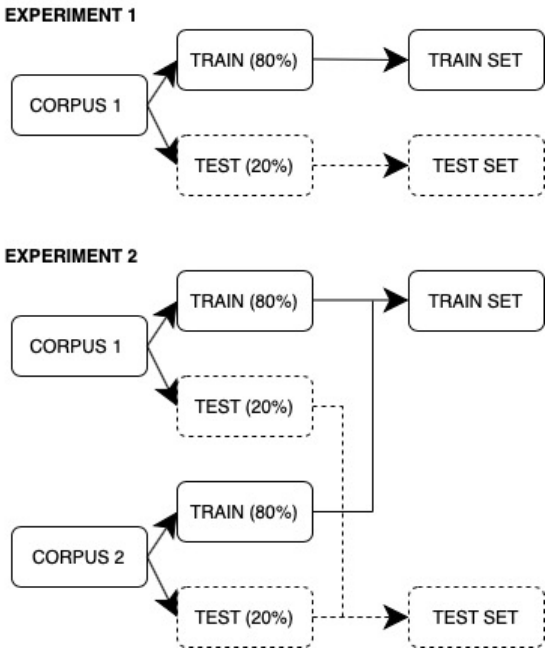



**FIGURE 3.** Schematic of the training/testing set creation procedure in Experiment 1 and Experiment 2.

subsets: 80% to be used during the training/validation phase and 20% to be used as the test set. We merged the two training groups into a single set, used to train the model with signals recorded both in controlled and domestic environment. Then, we also merged the two testing groups to carry out the testing phase on a collection of samples reflecting the characteristics of the training set. Given the limited numerosity of the second corpus, this procedure was preferred to using the non-supervised database only, so as to guarantee technically sound results. Indeed, although it provides only preliminary results on the possibility of using samples collected in unsupervised conditions, a comparison with Experiment 1 can provide pivotal information on the influence of the recording modality.

In Figure 3 we present a schematic of the training/testing set creation in Experiment 1 and Experiment 2. Also in this case, feature selection, model selection, and model optimization were performed on the training/validation set.

Taking advantage of having many different phonemes pronounced by the same subjects, we created the classifier input by merging into a single vector the features extracted from all the examined segments. Then, we compared the performance of 7 classifiers and optimized the one characterized by the highest accuracy. We tested classical approaches such as Naive Bayes (NB), k-Nearest Neighbor (KNN), SVM and RF, as well as ensemble methods such as Adaptive Boosting (ADA), Gradient Boosting (GB), and Bagging ensemble (BAG) classifiers. Given the random splitting procedure intrinsic in the validation process, we performed each experiment 20 times on 20 randomly extracted subsets, and considered the average accuracy as a suitable metric for comparison among classifiers.

After selecting the best classifier and optimizing its hyperparameters, we evaluated accuracy, f1 score, precision, and recall as an average of 20 iterations, to further assess the stability of the final model.

## IV. RESULTS AND DISCUSSION

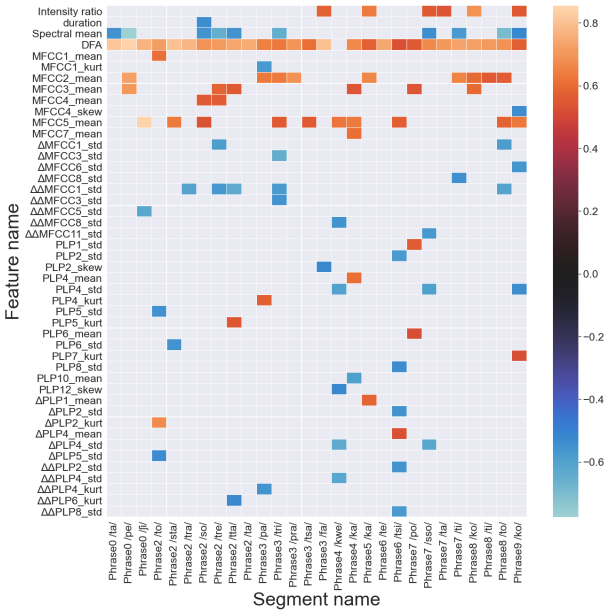
In this section, we present and discuss the results of the current study. We focus on the examination of the individual phonemes and the features extracted to verify if there are sounds of the Italian language more suitable for differentiating between PDPs and HCs, and if distinct sounds require different features to be described. Then, we report the classification results between PDPs and HCs and assess the influence of the recording conditions on the model.

### A. PHONETIC GROUPS EXAMINATION

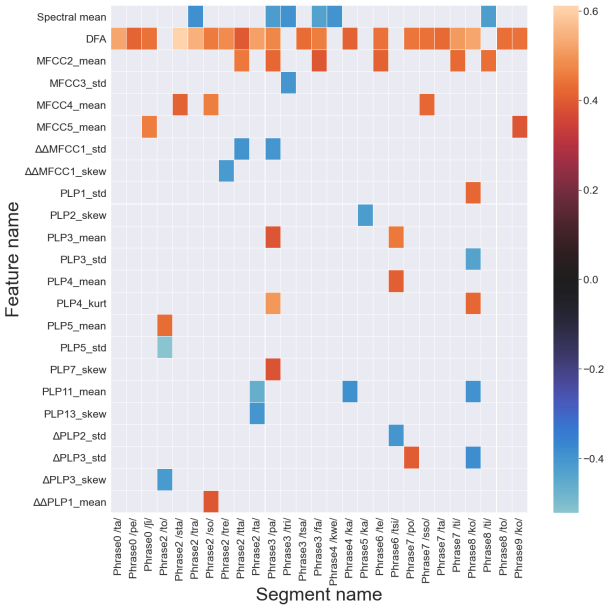
The segmentation of the ten sentences spoken by each subject identified 32 unique phonetic groups that included an unvoiced/voiced switch. Some of these sounds were pronounced multiple times in different sentences, thus the overall dataset consisted of a totality of 43 phonetic groups. Among these, 28 were correctly pronounced by all the individuals (i.e. no syllables missing or no word mispronounced); subsequent investigations focused on this subgroup. Based on the examination of the unvoiced consonant pronounced in the phonetic groups, it is possible to identify 13 dental occlusives, 5 velar occlusives, 4 labial occlusives, 2 alveolar sibilants, 1 palatal sibilant, 2 alveolar affricates, and 1 labio-dental fricative. Therefore, although there is unevenness between classes, it is possible to consider the employed dataset representative of the unvoiced consonants of the Italian language (i.e. dental occlusives, velar occlusives, labial occlusives, alveolar sibilants, palatal sibilant, alveolar affricates, palatal affricates, and labio-dental fricative).

In this section, we applied the feature selection described in section III-D to each phonetic group. Then, we calculated the Pearson correlation coefficient between the selected features and the class of membership, to identify the sounds and attributes with the highest discriminating capability. In Fig. 4 we report a schematic of the coefficients having a p value $< 0.001$. Fig. 4a puts into evidence the high potential of the TR between unvoiced consonants and the adjacent sound tract. Indeed, in Experiment 1, 28 phonetic segments exhibit at least two features with correlation to the membership class between 0.52 and 0.85 (absolute values). Among these, the DFA coefficient derived from the transition between the occlusive /p/ and the vowel /e/ and the fifth MFCC derived from the TR between the sibilant /ʃ/ and the vowel /i/ show the highest correlation with the class ($r = 0.85$, $p < 0.001$).

On the other hand, in Experiment 2 (Fig. 4b), although we can appreciate a general reduction of the performance ($0.37 < |r| < 0.62$), the results still reveal numerous features highly correlated to the membership class. Among these, the DFA derived from the transition between the occlusive consonant /t/ and the vowel /a/ shows the highest correlation with the class ($r = 0.62$, $p < 0.001$).




(a) *Experiment 1*




(b) *Experiment 2*

**FIGURE 4.** **Pearson correlation coefficient of the most relevant features and phonemes. Only those values associated with a p-value < 0.001 are reported.**

As for the specific types of features selected, the analysis puts into evidence that DFA, MFCC2, MFCC3, MFCC5, *intensity ratio*, and *spectral mean* exhibit a high correlation with the class for many of the selected phonetic groups when considering only signals recorded in optimal noise conditions. The introduction of samples belonging to the second dataset leads to a general reduction in the number of significant features. Indeed, Fig. 4a exhibits some differences with respect to Fig. 4b. In particular, it is worth noticing that the RASTA-PLP coefficients are the most selected features,

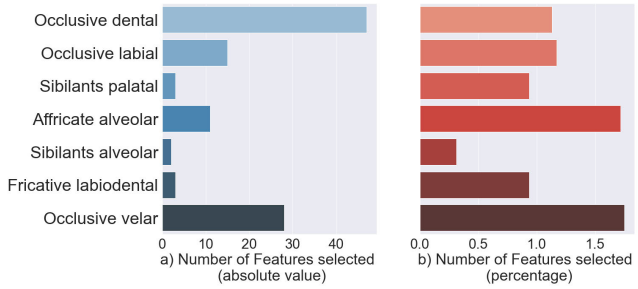



**FIGURE 5.** **Distribution of the features selected among phonetic groups - Experiment 1.**

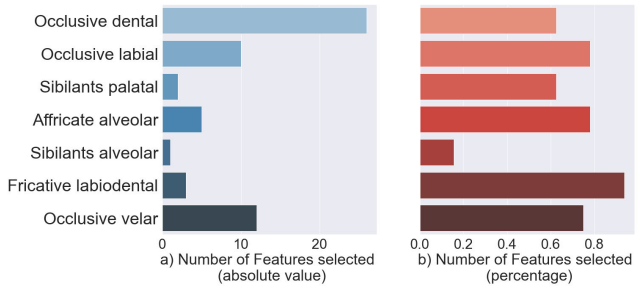



**FIGURE 6.** **Distribution of the features selected among phonetic groups - Experiment 2.**

suggesting an improved capability to capture differences between PDPs and HCs, even in the presence of sub-optimal quality records.

In a second section of our analysis, we investigated which sounds are the most representative of the language impairment, studying the features selected based on the phonetic group to which they refer.

Fig. 5 reports the results of the analysis conducted for Experiment 1. In particular, Fig. 5a shows the number of features selected in absolute value, whereas Fig. 5b takes into account the unevenness number of phonetic groups in the employed data sets, and represents the number of features selected in relation to the number of total segments belonging to each phonetic group.

In line with previous studies [2], the most serious pronunciation impairments occur in the occlusive consonants due to the movement necessary to produce such sounds. Indeed, unlike other consonants where there is no complete closure of the vocal trait, the sound of the occlusive consonants is produced when the air coming from the lungs meets an obstacle created by a sudden change of position of the articulators. The inability to produce fine and rapid movements makes it difficult for PDPs to produce these sounds. As far as concerns the place of articulation, the velar occlusive consonants, produced by withdrawing the tongue towards the soft palate, seem to have the highest discriminating power in Italian native speakers. Fig. 6 shows the results of the same analysis conducted for Experiment 2.

Although also in this case the features related to occlusive consonants are often selected, an important role is played by

**TABLE 3.** Classification accuracy comparison among 7 classifiers employing a 10-fold CV. Experiment 1.

| Model | Accuracy(%) | Precision(%) | Recall(%) | Specificity(%) | F1 score(%) | AUC |
|---|---|---|---|---|---|---|
| NB | 92 ± 2.6 | 95 ± 1.6 | 94 ± 3.1 | 91 ± 3.5 | 93 ± 2.6 | 0.98 ± 0.020 |
| KNN | 96 ± 2.3 | 95 ± 2.9 | 100 ± 0.0 | 92 ± 5.6 | 97 ± 1.7 | 0.99 ± 0.004 |
| **SVM** | **98 ± 1.2** | **98 ± 1.6** | **100 ± 0.0** | **97 ± 2.4** | **98 ± 1.1** | **1 ± 0.00** |
| ADA | 88 ± 4.8 | 90 ± 4.6 | 92 ± 4.0 | 83 ± 8.3 | 89 ± 4.3 | 0.92 ± 0.037 |
| GB | 87 ± 3.5 | 89 ± 3.5 | 93 ± 4.3 | 82 ± 6.9 | 88 ± 3.4 | 0.90 ± 0.051 |
| BAG | 97 ± 2.2 | 96 ± 2.6 | 100 ± 0.0 | 93 ± 5.4 | 98 ± 1.5 | 0.99 ± 0.008 |
| RF | 96 ± 1.8 | 96 ± 1.6 | 98 ± 2.3 | 93 ± 2.9 | 96 ± 1.8 | 0.99 ± 0.05 |

the parameters related to fricative labiodental that, in the current database, are represented by the syllable /fa/. However, additional analysis employing a larger dataset with a balanced distribution among phonetic groups is required to validate these findings.

### B. CLASSIFICATION RESULTS

In this section, we present the results of the binary classification between PDPs and HCs by mean of the procedure described in section III. In Table 3 we report the comparison of the classification accuracy on the 7 models tested.

The values refer to a 10-fold CV and are averaged over 20 iterations. As can be appreciated, the best results were achieved using the SVM classifier, which yields an accuracy of $98\% \pm 1.2$. It is worth noting that this latter classifier, besides achieving the highest accuracy, is also characterized by the lowest standard deviation, which is indicative of good stability of the model. Once identified the best classifier, we run a grid-search hyper-parameter optimization procedure within the training set, based on the mis-classification error minimization in 10-fold CV.

We tuned the C parameter in [10, 100, 1000] and gamma in [0.1, 0.001, 0.0001]. Moreover we investigated the performance achieved employing a linear, polynomial, and RBF kernels. The best configuration turned out to be an SVM with C = 10, gamma = 0.001, kernel = RBF. The results of the optimized SVM on the validation and test sets are reported in Table 4. As can be noticed from the analysis of the reported results, the performance does not impair when moving from validation data to completely new samples contained in the test set. This denotes the absence of over-fitting and a good generalization capability of the selected model.

Also, the phonetic groups employed and the types of features selected are reported in Table 4. As for the phonetic groups, the results trace the analysis conducted in section IV-A on the single phonemes and remark the importance of occlusive consonants as well as palatal sibilants.

Although there is no work in the state of art that lends itself to a fair comparison due to the different employed corpora, languages, and methods, the performance analysis of the most similar study employing TRs [1] revealed the high potential of the presented algorithm. In fact, as discussed in section II, in this latter work the authors achieved $94\% \pm 1$ accuracy (AUC = 0.99, Sens = 0.9, Spec = 1) in a 11-fold CV and $82\% \pm 13$ (AUC = 0.95, Sens = 1, Spec = 0.57) in the cross corpora experiments employing a GMM-UBM

**TABLE 4.** Performance of the optimized SVM model. Results are expressed as an average over 20 iterations. Experiment 1.

| Metric | Val. set | Test set | Phonetic group | Type of feat. |
|---|---|---|---|---|
| **Acc.(%)** | 98 ± 1.1 | 97 ± 5.6 | dental occlusive, labial occlusive, palatal sibilant, velar occlusive | DFA, spect. mean, MFCC2, MFCC5 |
| **Pre.(%)** | 98 ± 1.6 | 96 ± 7.4 | | |
| **Rec.(%)** | 99 ± 1.5 | 100 ± 0.0 | | |
| **Spec.(%)** | 97 ± 2.4 | 93 ± 13.9 | | |
| **F1(%)** | 98 ± 1.0 | 98 ± 4.1 | | |
| **AUC** | 1 ± 0.0 | 0.96 ± 0.07 | | |

classifier, PLP as features and the DDK speech task. Moreover, the accuracy reported when considering the same task considered in this work (i.e. text dependent utterance) is $89\% \pm 7$ (AUC = 0.93, Sens = 0.91, Spec = 0.91) in a 11-fold CV.

As far as concerns the classification performance obtained in Experiment 2, Table 5 reports the comparison of the classification accuracy on the 7 tested models. The values refer to a 10-fold CV and are averaged over 20 iterations. As can be noticed, the best results were achieved using an SVM classifier, which yields an accuracy of $87\% \pm 2.5$.

Again, once identified the best classifier, we run a grid-search hyper-parameters optimization procedure within the training set, based on the mis-classification error minimization in 10-fold CV. In more detail, we tuned the C parameter in [10, 100, 1000] and gamma in [0.1, 0.001, 0.0001]. Moreover we investigated the performance achieved employing a linear, polynomial, and RBF kernels. The best configuration turned out to be an SVM with C = 100, gamma = 0.001, kernel = RBF. The results of the optimized model are reported in Table 6.

As for the phonetic groups, the results remark the importance of occlusive and sibilants consonants. Furthermore, although features selected in Experiment 1 were mostly selected also in Experiment 2, in this latter case the set of attributes is larger, and includes RASTA-PLP coefficients and the derivatives of both RASTA-PLP and MFCC. Also in this second case, the performance remains stable when moving from validation to the test set, although the standard deviation slightly increases in the latter case. In figure 7 we report the ROC curve of the final model averaged over 20 iterations. A comparison between the two Experiments reveals that the set of features selected in Experiment 2 is a super-set of those selected in Experiment 1.

Therefore, we can assume that this enlarged feature ensemble is able to effectively capture vocal alterations in

**TABLE 5.** Classification accuracy comparison among 7 classifiers employing a 10-fold CV. Experiment 2.

| Model | Accuracy(%) | Precison(%) | Recall(%) | Specificity(%) | F1 Score(%) | AUC |
|---|---|---|---|---|---|---|
| NB | 83 ± 2.2 | 89 ± 2.5 | 82 ± 3.5 | 85 ± 3.5 | 84 ± 2.4 | 89 ± 2.7 |
| KNN | 84 ± 2.7 | 87 ± 3.7 | 85 ± 2.6 | 82 ± 4.5 | 85 ± 2.5 | 92 ± 2.4 |
| **SVM** | **86 ± 2.5** | **90 ± 2.9** | **87 ± 2.7** | **86 ± 3.9** | **87 ± 2.3** | **94 ± 1.9** |
| ADA | 84 ± 3.6 | 88 ± 3.6 | 85 ± 3.4 | 83 ± 5.2 | 85 ± 3.2 | 91 ± 3.8 |
| GB | 76 ± 3.5 | 82 ± 4.0 | 77 ± 4.4 | 75 ± 6.0 | 77 ± 3.6 | 83 ± 3.7 |
| BAG | 84 ± 2.8 | 88 ± 3.4 | 85 ± 3.2 | 83 ± 4.2 | 85 ± 2.8 | 93 ± 2.2 |
| RF | 83 ± 2.9 | 87 ± 3.0 | 84 ± 3.4 | 82 ± 3.5 | 84 ± 3.0 | 93 ± 2.8 |

**TABLE 6.** Performance of the optimized SVM model. Results are expressed as an average over 20 iterations. Experiment 2.

| Metric | Val. set | Test set | Phonetic group | Type of feat. |
|---|---|---|---|---|
| **Acc.(%)** | 88 ± 2.8 | 90 ± 6.8 | labiodental fricative; dental, labial, and velar occlusives; affricate and sibilants alveolars; palatal sibilants | DFA, spect. mean, MFCC4, ΔMFCC1, ΔΔMFCC3, PLP3, PLP5, PLP11 ΔPLP3, ΔΔPLP1 |
| **Pre.(%)** | 93 ± 3.0 | 95 ± 5.0 | | |
| **Rec.(%)** | 87 ± 2.9 | 88 ± 10.0 | | |
| **F1(%)** | 89 ± 2.7 | 91 ± 6.3 | | |
| **Spec.(%)** | 89 ± 4.1 | 93 ± 7.4 | | |
| **AUC** | 94 ± 2.4 | 91 ± 6.3 | | |

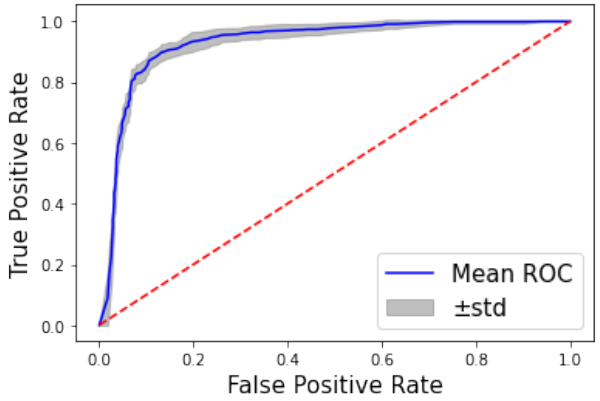



**FIGURE 7.** ROC curve expressed as an average over 20 iterations. Experiment 2.

sub-optimal recordings. It is also worth noticing that the inclusion of non-supervised recordings leads to substantial changes to the model encompassing both the features and the hyperparameters selected. This suggests that the algorithm is able to efficiently adapt to the input data. Hence, one can expect that, if the model is trained on a larger dataset, it can lead to optimal performance regardless of the sub-optimal recording conditions.

## V. CONCLUSION AND FUTURE WORK

In this study, we investigated the impact of PD on unvoiced/voiced transitions and the discriminatory capacity of different phonemes in the Italian language. To this aim, we employed features already involved in similar analysis in conjunction with novel parameters.

This work has confirmed the possibility of a speech-based PD classification model and the effectiveness of the TR analysis. Furthermore, the investigation of the discriminatory capability of various Italian phonemes confirmed that the most critical pronunciation problems occur in occlusive consonants. As far as concerns the place of articulation, the velar occlusive consonants, produced by withdrawing the tongue towards the soft palate, seem to have the highest discriminant power in Italian native speakers.

Although these results can provide a powerful tool for the analysis of Italian PD speakers, the heterogeneity of the used dataset requires further investigations to validate these preliminary findings. Moreover, despite we are aware that the need for a manual segmentation represents a limit for the work, in this stage of our research we wanted to avoid to bias the results using an automatic segmentation tool. Actually, such tools are affected by an intrinsic source of error, even more evident in the case of PDPs, whose speech is severely impaired. Once the effectiveness of using transition zones is fully validated, we plan to address an automatic segmentation tool to ease the segmentation procedure and extend the results to broader corpora and languages. Furthermore, we plan to extend the model developed in this paper to other languages, in order to perform comparisons and identify similarities and differences.

As for the types of features employed, our results suggest the high potential of the DFA coefficient, which, to the best of our knowledge, has never been applied to the TR analysis.

Finally, this work discloses the possibility of voice processing employing recordings collected both in optimal and sub-optimal conditions. Indeed, we performed the collection of an additional database through a web application that guided the users through the execution of the same tasks encompassed in the main dataset. If properly validated, this data collection technique would enable a frequent monitoring of the disease. The recordings could be performed in a comfortable environment, so obtaining voice samples that reflect more closely the actual condition of the patient. This study is part of an ongoing project to develop a lightweight system that can be employed for the home monitoring of patients.

However, the small numerosity of the two corpora and the slight difference in the average age of the two groups in the second dataset prevents from considering these results as exhaustive; hence we plan to perform further validation with a much larger cohort of subjects

Besides collecting more speech data from PDPs, we also plan to employ clinical information as the H&Y stage and UPDRS-Part II/Part III scores, and to perform data acquisition several times on the same patients, both under and not under dopaminergic therapy, to verify whether the analysis of the TR applies to these fields also.

Specifically, the evidence from this study suggests the feasibility of a tool that may be employed for home monitoring of motor fluctuations in PDPs, as well as for early PD diagnosis decision support. On the other hand, given the reduced size of the dataset employed, our methods and results require further validation with a much larger cohort of subjects.

At last, this is part of a more large PD monitoring study [29], which we plan to implement in a sort of electronic diary of PDPs, including motor symptoms and postural control monitoring, as well as sleep quality assessment.

## ACKNOWLEDGMENT

The authors would like to thank the Associazione Amici Parkinsoniani Piemonte Onlus and the Imperia Hospital (Neurology Department) for their contribution to the data collection procedure.

## ACRONYMS

| | |
|---|---|
| ADA | Adaptive Boosting |
| AUC | Area Under the Curve |
| BAG | Bagging ensemble |
| CV | Cross Validation |
| DDK | Diadochokinetic |
| DFA | Detrended fluctuation analysis |
| ETS | Energy Transition Slope |
| GB | Gradient Boosting |
| GMM | Gaussian Mixture Models |
| H&Y | Hoehn and Yahr |
| HC | Healthy Control |
| KNN | k-Nearest Neighbor |
| LOO | Leave One Out |
| MDS | Movement Disorder Society |
| MFCC | Mel Frequency Cepstral Coefficients |
| ML | Machine Learning |
| NB | Naive Bayes |
| PD | Parkinson's Disease |
| PDP | Patients with Parkinson's Disease |
| PLP | Perceptual Linear Prediction |
| RASTA-PLP | Relative Spectral - Perceptual Linear Prediction |
| RBF | Radial Basis Function |
| RF | Random Forest |
| ROC | Receiving Operator Curve |
| SVM | Support Vector Machine |
| TR | Transition Regions |
| UBM | Universal Background Models |
| UPDRS | Unified Parkinson's Disease Rating Scale |
| VOT | Voice Onset Time |

## REFERENCES

[1] L. Moro-Velazquez, J. Gomez-Garcia, N. Dehak, and J. I. Godino-Llorente, "New tools for the differential evaluation of Parkinson's disease using voice and speech processing," in *Proc. IberSPEECH*, 2021, pp. 165–169.

[2] J. R. Orozco-Arroyave, *Analysis of Speech of People With Parkinson's Disease*. Logos Verlag Berlin GmbH, 2016, p. 138.

[3] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol., Neurosurg. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[4] P. Gómez-Vilda, J. Mekyska, J. M. Ferrández, D. Palacios-Alonso, A. Gómez-Rodellar, V. Rodellar-Biarge, Z. Galaz, Z. Smekal, I. Eliasova, M. Kostalova, and I. Rektorova, "Parkinson disease detection from speech articulation neuromechanics," *Frontiers Neuroinform.*, vol. 11, p. 56, Aug. 2017.

[5] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art," *Biomed. Signal Process. Control*, vol. 51, pp. 181–199, May 2019.

[6] C. Manfredi, M. D'Aniello, P. Bruscaglioni, and A. Ismaelli, "A comparative analysis of fundamental frequency estimation methods with application to pathological voices," *Med. Eng. Phys.*, vol. 22, pp. 135–147, Mar. 2000.

[7] J. Rusz, J. Hlavnicka, T. Tykalova, M. Novotny, P. Dusek, K. Sonka, and E. Ruzicka, "Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1495–1507, Aug. 2018.

[8] H. C. Tunc, C. O. Sakar, H. Apaydin, G. Serbes, A. Gunduz, M. Tutuncu, and F. Gurgen, "Estimation of Parkinson's disease severity using speech features and extreme gradient boosting," *Med. Biol. Eng. Comput.*, vol. 58, no. 11, pp. 2757–2773, 2020.

[9] R. B. Postuma, A. E. Lang, J. F. Gagnon, A. Pelletier, and J. Y. Montplaisir, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

[10] J. Hlavnika, R. Cmejla, T. Tykalová, K. Šonka, E. Ruzicka, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Sci. Rep.*, vol. 7, no. 1, p. 12, 2017.

[11] A. Ma, K. K. Lau, and D. Thyagarajan, "Voice changes in Parkinson's disease: What are they telling us?" *J. Clin. Neurosci.*, vol. 72, pp. 1–7, Feb. 2020, doi: 10.1016/j.jocn.2019.12.029.

[12] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, "Advances in Parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102418.

[13] J. Jacobi, T. Rebernik, R. Jonkers, M. Proctor, B. Maassen, and M. Wieling, "The effect of Levodopa on vowel articulation in parkinson's disease: A cross-linguistic study," in *Proc. 19th ICPhS*, 2019, pp. 1069–1073.

[14] J. C. Vásquez-Correa, T. Arias-Vergara, C. D. Rios-Urrego, M. Schuster, J. Rusz, S. Shattuck-Hufnagel, and E. Nöth, "Convolutional neural networks and a transfer learning strategy to classify Parkinson's disease from speech in three different languages," in *Proc. Iberoamerican Congr. Pattern Recognit.*, in Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11896. Cham, Switzerland: Springer, 2019, pp. 697–706.

[15] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, and N. Dehak, "A forced Gaussians based methodology for the differential evaluation of Parkinson's disease by means of speech processing," *Biomed. Signal Process. Control*, vol. 48, pp. 205–220, Feb. 2019, doi: 10.1016/j.bspc.2018.10.020.

[16] J. I. Godino-Llorente, S. Shattuck-Hufnagel, J. Y. Choi, L. Moro-Velázquez, and J. A. Gómez-García, "Towards the identification of idiopathic Parkinson's disease from the speech. New articulatory kinetic biomarkers," *PLoS ONE*, vol. 12, no. 12, 2017, Art. no. e0189583.

[17] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, F. Grandas-Perez, S. Shattuck-Hufnagel, V. Yagüe-Jimenez, and N. Dehak, "Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's disease," *Sci. Rep.*, vol. 9, no. 1, p. 19066, 2019.

[18] K. Chenausky, J. MacAuslan, and R. Goldhor, "Acoustic analysis of PD speech," *Parkinson's Disease*, vol. 2011, pp. 1–13, Jan. 2011.

[19] Y. Kim, "Acoustic characteristics of fricatives /s/ and /ʃ/ produced by speakers with Parkinson's disease," *Clin. Arch. Commun. Disorders*, vol. 2, no. 1, pp. 7–14, 2017.

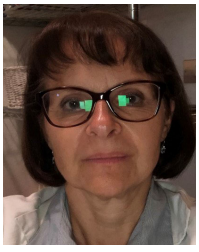
[20] A. Hernandez and M. Chung, "Dysarthria classification using acoustic properties of fricatives," in *Proc. Seoul Int. Conf. Speech Sci. (SICSS)*, 2019, pp. 43–44.

[21] R. Viswanathan, P. Khojasteh, B. Aliahmad, S. P. Arjunan, S. Ragnav, P. Kempster, K. Wong, J. Nagao, and D. K. Kumar, "Efficiency of voice features based on consonant for detection of Parkinson's disease," in *Proc. IEEE Life Sci. Conf. (LSC)*, Oct. 2018, pp. 49–52.

[22] D. Montaña, Y. Campos-Roca, and C. J. Pérez, "A Diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson's disease," *Comput. Method Programs Biomed.*, vol. 154, pp. 89–97, Feb. 2018.
[23] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22199–22208, 2017.
[24] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of Parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019.
[25] D. Hemmerling and M. Wojcik-Pedziwiatr, "Prediction and estimation of Parkinson's disease severity based on voice signal," *J. Voice*, vol. S0892-1997, no. 20, pp. 30231–30239, Aug. 2020.
[26] S. Arora, L. Baghai-Ravary, and A. Tsanas, "Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice," *J. Acoust. Soc. Amer.*, vol. 145, no. 5, pp. 2871–2884, May 2019, doi: 10.1121/1.5100272.
[27] B. Karan, S. S. Sahu, J. R. Orozco-Arroyave, and K. Mahto, "Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech," *Biomed. Signal Process. Control*, vol. 61, Aug. 2020, Art. no. 102050.
[28] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010.
[29] L. Borzì, M. Varrecchia, G. Olmo, C. A. Artusi, M. Fabbri, M. G. Rizzone, A. Romagnolo, M. Zibetti, and L. Lopiano, "Home monitoring of motor fluctuations in Parkinson's disease patients," *J. Reliab. Intell. Environ.*, vol. 5, no. 3, pp. 145–162, 2019.

**FEDERICA AMATO** received the B.S. and M.Sc. degrees in biomedical engineering from the Politecnico di Torino, Italy, in 2017 and 2020, respectively, where she is currently pursuing the Ph.D. degree with the Department of Control and Computer Engineering. From 2018 to 2019, she studied at the University of Cork, Ireland. Her M.Sc. degree discussing a thesis on the automatic assessment of Parkinson's Disease (PD) by means of speech analysis. Her research interests include speech processing with a particular focus on early diagnosis of vocal disturbances in PD patients. Her research studies focus on automated speech analysis for remote PD monitoring and early diagnosis of speech disturbances.

**LUIGI BORZÌ** (Member, IEEE) received the B.S. and M.Sc. degrees in biomedical engineering from the Politecnico di Torino, in 2015 and 2018, respectively, where he is currently pursuing the Ph.D. degree in computer engineering. In 2018, he was a Research Assistant at the Department of Control and Computer Engineering, Polytechnic University of Turin, and the Department of Neuroscience, "Città della Salute e della Scienza," Turin. His research studies focus on the remote monitoring of PD clinical condition using wearable sensors and artificial intelligence. His research interests include remote monitoring of PD motor symptoms, automated postural stability and fall risk assessment, and early diagnosis of speech disturbances and sleep disorders.



**GABRIELLA OLMO** (Senior Member, IEEE) received the Ph.D. degree *(summa cum laude)* in electronic engineering from the Politecnico di Torino, in 1992. She graduated in medicine and surgery from the University of Turin, in 2016, discussing a thesis on microtubule instability. She was an Assistant Professor and then an Associate Professor at the Department of Electronics, Politecnico di Torino, where she founded a research group on image/video coding and multimedia . In 2017, she moved at the Department of Computer Science, Politecnico di Torino. She collaborates with the Molecular Biotechnology Center, University of Turin, the Center for Neuroscience and Cognitive Systems, Italian Institute of Technology, and the Center for Parkinson's Disease and Movement Disorders, "City of Health and Science" Hospital, Turin. She has coauthored more than 200 peer-reviewed articles. Among her multidisciplinary interests, signal processing and machine learning for neurodegenerative diseases. She is a member of the Italian Society of Neurology and the Movement Disorders Society.

**CARLO ALBERTO ARTUSI** graduated in medicine and surgery from the University of Turin, in 2012, with a thesis on deep brain stimulation. He is currently pursuing the Ph.D. degree with the Doctoral School in Neuroscience, University of Torino. In June 2010 and 2013, he attended the Neurology Unit, University of Turin, where he did research on Parkinson's disease. In Summer 2017, he was at the Gardner Family Center for Parkinson's Disease and Movement Disorders, Department of Neurology, University of Cincinnati, Cincinnati, OH, USA. From 2013 to 2018, he attended the School of Specialization in Neurology, University of Torino. He is currently working as a Consultant Neurologist at the Fondazione Don Carlo Gnocchi and Humanitas–Gradenigo Hospital, Torino, Italy. His activities were in the field of Parkinson's disease and demyelinating disorders. His clinical activity is in the field of neurodegenerative diseases, with particular focus on advanced therapies for Parkinson's disease.

**GABRIELE IMBALZANO** graduated in medicine and surgery from the University of Messina, in 2016, carrying out an experimental thesis entitled "Disturbi cognitivi e psichiatrici inpazienti parkinsoniani sottoposti a stimolazione cerebrale profonda." In 2017, he received the Italian Habilitation as a Medical Doctor. He worked as a Temporary Substitute of General Practitioner at Azienda Sanitaria Provinciale (ASP) of Reggio Calabria. Since December 2017, he has been a Neurology Resident at the Department of Neuroscience "Rita Levi Montalcini," University of Turin. His clinical activity and scientific research interests include clinical neurology, particularly clinical management and research of Parkinson's disease and other movement disorders.

**LEONARDO LOPIANO** graduated in medicine and surgery from the University of Turin. He received the Ph.D. degree in neurological sciences from the University of Turin, in 1989, and specialization in neurology, in 1992. From 1994 to 2002, he was an Assistant Professor and the Medical Director at the University Division "Neurology I," Turin. From 2002 to 2005, he was an Associate Professor at the Department of Neuroscience, University of Turin. Since 2005, he has been a Full Professor with the Department of Neuroscience, University of Turin. He is currently the Director of the University Division "Neurology 2," "City of Health and Science" Hospital, Turin. He was a Coordinator of the Study Group "Cerebral Deep Stimulation," Italian Society of Neurology (SIN), a Secretary and a Treasurer of the Italian league for the fight against Parkinson's Disease, Extrapyramidal Syndromes and Dementia (LIMPE), and an Executive Council Member of the DISMOV Association. He is a member of the Board of Directors of SIN and the President of the LIMPE/DISMOV Academy.