Mittapalle, Kiran; Alku, Paavo

Exemplar-based Sparse Representations for Detection of Parkinson's Disease from Speech

Published in:
IEEE/ACM Transactions on Audio Speech and Language Processing

DOI:
10.1109/TASLP.2023.3260709

Published: 27/03/2023

Document Version
Publisher's PDF, also known as Version of record

Exemplar-Based Sparse Representations for Detection of Parkinson's Disease From Speech

Mittapalle Kiran Reddy and Paavo Alku, *Fellow, IEEE*

***Abstract*—Parkinson's disease (PD) is a progressive neurological disorder which affects the motor system. The automatic detection of PD improves the diagnosis of the disease, and it can be done in a non-invasive manner from speech. In this paper, we investigate the use of an exemplar-based sparse representation (SR) classification approach for detecting PD from speech. Exemplars are speech feature vectors extracted from the training data. The idea is to formulate the detection task as a problem of finding sparse representations of test speech feature vectors with respect to training speech exemplars. The main advantage of using the SR approach instead of conventional machine learning (ML)-based approaches is that the training step–which is time-consuming and sometimes requires unorganized hyper-parameter tuning–is not needed. Furthermore, SRs are more robust to redundancy and noise in the data. In this work, we study SR classification approaches based on two sparse coding models, namely, l1-regularized least squares (l_1LS) and non-negative least squares (NNLS). We propose a strategy based on class-specific dictionaries for improving performance of the l_1LS- and NNLS-based SR classification. To investigate the detection performance, the l_1LS- and NNLS-based approaches are applied and compared with the traditional PD detection approach based on ML classification algorithms using the PC-GITA PD dataset and an openly available dataset consisting of mobile device voice recordings from healthy and PD patients. The results indicate that the proposed NNLS-based SR classification approach performs better than the traditional ML-based methods in discriminating PD patients from healthy subjects.**

***Index Terms*—Exemplar-based, glottal features, Parkinson's disease, non-negative least squares, random forest, sparse representation, SVM.**

I. Introduction

PARKINSON'S disease (PD), initially called shaking palsy, is a neuro-degenerative disease caused by damage to the dopaminergic neurons in the mid-brain region [1], [2], [3]. The dopaminergic neurons play an important role in controlling multiple brain functions in voluntary movements of the muscles in the face and mouth to generate speech [2], [3]. The lower the level of dopamine, the higher the probability of being affected by PD [2]. Studies have shown that PD primarily causes speech deficits at the early stages of the disease. Speech disorders caused by PD can be characterized by symptoms such as reduced tongue flexibility and reduced vocal tract volume, reduced speech intensity and pitch range, impairments in speech quality, and inappropriate pauses [4]. Therefore, the detection of PD, particularly in its incipient stages, from the speech signal is a justified and an important topic [4], [5].

In the literature of automatic PD detection, most of the works make use of the classical pipeline approach, which consists of two separate stages (feature extraction and classification). As an alternative to the classical pipeline approach, some studies have recently investigated the PD detection task using the end-to-end deep learning framework consisting of convolutional neural networks (CNNs) and multilayer perceptrons (MLPs) [26], [27], [30]. End-to-end systems are completely data-driven and do not require any domain expertise in PD [26]. However, large amounts of data are required to properly train deep learning models. Collecting large amounts of training data is, however, difficult from PD patients because they might not bear long recordings. Due to this data sparsity issue, classical pipeline systems are still a justified choice for automatic detection of PD from speech and therefore the focus of the current study is also on these systems. A summary of previous studies in the detection of PD from speech signals based on the classical pipeline approach is provided in the following section. For a complete review on speech signal processing algorithms for PD detection, the reader is referred to [6], [9].

A. Related Works

In the classical pipeline approach, a machine learning (ML) classifier is trained with a discriminative set of hand-crafted speech features to identify individuals with PD. The features are mainly used to model the two main aspects of speech signals: articulation and phonation. As described in [6], the commonly used phonation features include, for instance, jitter, shimmer, harmonic-to-noise ratio (HNR), noise-to-harmonic ratio (NHR), pitch, and Mel-frequency cepstral coefficients (MFCCs). The most popular articulation features include MFCCs, features based on linear predictive coding (LPC), and perceptual linear prediction (PLP). In [12], a support vector machine (SVM) classifier was trained using a set of selected traditional phonation measures (such as jitter and shimmer) and pitch period entropy extracted from sustained phonations for the discrimination of PD from healthy subjects. In [13], the authors compared the performance of the k-nearest neighbors (k-NN) and SVM methods in classifying subjects with PD. The classifiers were

Manuscript received 12 July 2022; revised 9 January 2023 and 3 March 2023; accepted 8 March 2023. Date of publication 27 March 2023; date of current version 7 April 2023. This work was supported by the Academy of Finland under Grant 330139. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Juan Ignacio Godino-Llorente. *(Corresponding author: Mittapalle Kiran Reddy.)*

The authors are with the Department of Signal Processing and Acoustics, Aalto University, 00076 Espoo, Finland (e-mail: kiran.r.mittapalle@aalto.fi; paavo.alku@aalto.fi).

Digital Object Identifier 10.1109/TASLP.2023.3260709

trained using a group of 26 linear and time-frequency-based speech features. The results showed that the SVM provided more stable results compared to the k-NN classifier. A method based on 132 dysphonia measures computed from sustained vowels was proposed for PD detection in [14]. In their study, subsets of dysphonia measures obtained via feature selection were used to train SVM and random forest (RF) classifiers, and SVM was shown to perform better than RF in detecting speakers with PD. In [15], the authors proposed an approach based on a genetic algorithm (GA) and SVM for the discrimination of PD patients from healthy controls. The GA was used to select optimal speech features from a set of 22 linear and non-linear features. The selected features were used to train the SVM classifier for the PD detection task. In [18], a decision-tree-based ensemble ML algorithm was studied to predict the progress of PD based on phonatory features. The use of the integrated chaotic bacterial foraging optimization (CBFO) with an improved fuzzy k-NN classifier was studied in early diagnosis of PD using 22 vocal features in [19]. The authors reported that the CBFO-FKNN approach performed better than SVM in the diagnosis of PD using sustained vowels.

MFCCs have been widely used as robust articulation features in PD classfication [6], [7], [9]. In [8], MFCCs were applied in the classification of speech by PD patients and healthy controls (HCs) using the Spanish PC-GITA database [50]. The authors trained SVMs with statistical functionals of MFCCs, which were computed on the Bark bands from speech that was collected using different speech tasks (reading individual sentences, text reading, the diadochokinetic (DDK) task, and monologue). As a result, an accuracy of 73.7% was reported based on K-fold (K $=$ 10) cross-validation [8]. In [20], MFCCs were used in analysis of continuous speech produced by speakers with PD by studying recordings made in different languages (Spanish, German, and Czech). In [22], MFCCs along with other cepstral features such as PLP features were analyzed in the discrimination of healthy individuals from PD patients using voice recordings of the vowel /a/. The method in [20] was based on modeling the energy content of unvoiced sounds using MFCCs and Bark bands energies that were extracted from speech produced in various speech tasks (isolated words and sentences, the DDK and text reading tasks). In [9], MFCCs were shown to perform comparably to or better than many other features such as perturbation measures [6], complexity measures [7] or measures based on the tunable Q-factor wavelet transform, for utterances of the vowel /a/. Recently, in [21], SVMs trained with cepstral coefficients derived using the single frequency filtering (SFF)-based instantaneous spectral representation was studied for PD detection. The authors conducted a comparative study on the PD detection task using sustained vowels and reported that SFF-based features perform better than MFCCs. However, the major disadvantage of the SFF-based features is the high complexity involved in the extraction of these features. In [23], the authors trained an SVM classifier by combining formant features extracted with Hilbert-Huang Transform and MFCCs. It was observed that the combination improved the PD detection performance for the vowels of the PC-GITA database. In [7], diagnosis of PD from speech was performed using articulation and perturbation measures based on regularization techniques. In [25], a larger feature set consisting of articulation, phonation and prosody features extracted from monologue sentences of the PC-GITA dataset was employed to develop an SVM model for classification of patients with PD. The authors observed that the combined features provided better performance compared to the individual features alone. In [26], an improvement in PD detection performance was observed when glottal source features were further combined with the larger feature set described in [25].

B. Scope of the Study

In classical pipeline systems, a learning-based classifier (such as SVM) is first trained using the extracted features in the training stage and then evaluated in a separate testing stage [31]. The performance of the ML classifier evaluated in the testing stage, however, depends significantly on how the hyper-parameters of the classifier are adjusted in the training stage, and the detection performance is sensitive to redundancy in the data [31], [32]. As an alternative to the widely used ML-based approaches, $exemplar-based$ sparse representations (SRs) have been used for improving the classification performance in some fields such as bio-informatics [31], face recognition [34], music genre classification [36] and speech recognition [35]. In the exemplar-based approach, speech exemplars from the training set are first collected into a dictionary. Then the sparse coefficient vector corresponding to a test instance is obtained using the dictionary [31], [32]. Finally, a class label can be predicted from the estimated sparse vector using a sparse interpreter [31], [32], [35]. The main advantage of the SR approach is that the classification can be performed $directly$ by representing the test sample as a sparse linear combination of the labeled training samples [31], [32]. Hence, the tedious training step, which includes unorganized hyper-parameter tuning, can be avoided. Furthermore, SR offers other advantages, such as strong robustness and less sensitivity to the selected features. Motivated by these issues, the current study will investigate SR in classification of speech either as parkinsonian or healthy and compare the performance of SR with reference ML-based approaches. To the best of our knowledge, the exemplar-based SR approach has not yet been considered for PD (or any other pathology) detection from speech signals.

Although several studies have investigated the use of phonation and articulation information in detection of parkinsonian speech demonstrating the relevance of both aspects (see [6] for a review), only a few studies have investigated the joint use of both aspects ([7], [25], [26]). These studies have demonstrated that the ML models developed with the combined features (articulation and phonation) provide better discrimination between healthy and PD speech. However, in examining the joint use of articulation and phonation features for PD detection, these studies have either considered data from only one speech task (e.g., vowels [7], monologue sentences [25]) or used data from different speech tasks together in building the classification models (eg., [26]). Furthermore, the effectiveness of the joint features in the automatic detection of PD using speech recorded by

mobile devices–data which is more suitable for telemonitoring applications–has not been studied much previously. Therefore, by combining our motivation to study the examplar-based SR approaches in PD detection with the above-described issues, the goal of the current study is summarized as follows: We compare exemplar-based SR approaches to traditional ML-based approaches in the automatic detection of PD from speech by using combined phonation and articulation features, by using different speech tasks, and by using speech datasets recorded in different conditions.

The organization of this paper is as follows. The exemplar-based SR approaches for classification of healthy and PD patients are described in Section II. The databases, experimental setup, and evaluation metrics are discussed in Section III. The experimental results are presented in Section IV. Finally, the conclusions are presented in Section V.

II. Exemplar-Based Sparse Representation Classification

SR is a parsimonious principle according to which a sample can be sparsely represented with a redundant dictionary of non-orthogonal basis vectors [31]. Given an n-dimensional input data $\mathbf{y}$, the SR problem is formulated as $\mathbf{y} \approx \mathbf{Dx}$, where $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k]$ is known as the dictionary and $\mathbf{d}_k$ is a dictionary atom (basis vector) and $\mathbf{x}$ is the sparse coefficient vector. Here, n denotes the dimension of the input test vector and k denotes the number of atoms in the dictionary. Sparse representation includes *sparse coding* and *dictionary learning* [32]. Given a new data and the dictionary $\mathbf{D}$, learning the sparse coefficient vector $\mathbf{x}$ is a procedure of sparse coding. Given training data, learning the dictionary from data is called dictionary learning. An exemplar-based SR approach falls under the category of sparse coding as it uses a fixed dictionary consisting of exemplars from the training set as basis vectors to determine $\mathbf{x}$. The sparse coefficient vector $\mathbf{x}$ can be obtained by solving the following problem

$$\mathbf{x} = \min_{\mathbf{x}'} \|\mathbf{x}'\|_0 \quad \text{s.t} \quad \mathbf{y} = \mathbf{Dx}. \tag{1}$$

Equation (1) is not convex and is also a non-deterministic polynomial (NP)-hard problem [41]. An NP-hard problem cannot be solved in polynomial time on a standard computer. Alternatively, the sparse coefficient vector $\mathbf{x}$ can be obtained by solving the following l_1-regularized least squares problem [31]

$$J(\mathbf{x}, \lambda) = \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{Dx}\|_2^2 + \lambda \|\mathbf{x}\|_1 . \tag{2}$$

Here, $\lambda > 0$ is the scalar regularization parameter that balances the trade-off between reconstruction error and sparsity. (2) is known as the l_1-least squares (l_1-LS) sparse coding model which can be solved efficiently using techniques like the truncated Newton interior-point method (TNIPM) proposed in [33]. The l_1-norm minimization can efficiently recover sparse signals [33] and is robust with respect to outliers. However, there are two major drawbacks with the l_1-LS model. First, it is a parametric approach, in which the value of λ needs to be properly chosen depending upon the application. Second, it results in a sparse coefficient vector consisting of mixed signs which are hard to interpret [32].

Algorithm 1: l_1-LS or NNLS sparse coding-based classification.

Input: $\mathbf{D}_{N\times M}$: training data including N-dimensional feature vectors and M samples, $\mathbf{c}$: class labels of the training samples, $\mathbf{Y}_{N\times p}$: p new samples (i.e. test samples)
Output: $\mathbf{p}$: predicted class labels of the p test samples
1) Learn the sparse coefficient vector $\mathbf{x}$, of each test sample by solving (3) or (4);
2) Predict the class labels of the test feature vector by using a sparse interpreter (or rule).

Instead of l_1-LS, the non-negative least-squares (NNLS) sparse coding model has been used for deriving sparse representations in [32], [40]. In the NNLS approach, the sparse vector $\mathbf{x}$ is obtained by solving the following equation

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{Dx}\|_2^2 \quad \text{s.t } \mathbf{x} \geq 0. \tag{3}$$

The active-set NNLS sparse coding algorithm [32], [42] is used for solving (4). There are three advantages of NNLS over l_1-LS. First, NNLS is a non-parametric model, which is more convenient in practice. Second, unlike in l_1-LS, the vector $\mathbf{x}$ is not allowed to contain negative values (as can be seen from (5)) with the NNLS model. In classification task, sparse interpreters (like the nearest-subspace (NS) rule discussed later in the paper) have been observed to work well with a non-negative sparse vector compared to mixed-sign vector [32], [40]. Third, unlike l_1-LS, the NNLS problems can be solved in batch, which makes the NNLS approach run much faster than its l_1-LS counterpart [32].

The sparse vector $\mathbf{x}$ generated with a sparse coding model can be interpreted directly to determine the class label, which is known as SR classification [31], [32]. The general steps in the exemplar-based SR classification approach using either the l_1-LS- or NNLS-based sparse coding model are detailed in Algorithm 1 [31], [32], [40]. First, the N-dimensional feature vectors extracted from the training samples are arranged as columns in dictionary $\mathbf{D}$. Then, either (3) or (4) is solved to derive sparse coefficient vector $\mathbf{x}$ corresponding to test feature vector $\mathbf{y}$. Finally, the class label of each new sample is predicted using a sparse interpreter. The MAX rule and nearest-subspace (NS) rule are the two popular sparse interpreters [34], [40]. However, when the sparsity decreases in signals either due to their inherent nature (like in the case of speech signals) or due to the presence of noise, the MAX rule may deliver incorrect decisions [40]. Therefore, in this work, we employed the nearest-subspace (NS) rule [34], which is more robust than the MAX rule when signals are less sparse. In the NS rule, suppose there are C classes with labels $l_1, ..., l_C$, after obtaining non-negative coefficient vector $\mathbf{x}$ corresponding to new sample $\mathbf{y}$, first the regression residual corresponding to the ith class is computed as

$$r_i(\mathbf{y}) = \frac{1}{2} \|\mathbf{y} - \mathbf{D}\delta_i(\mathbf{x})\|_2^2 \tag{4}$$

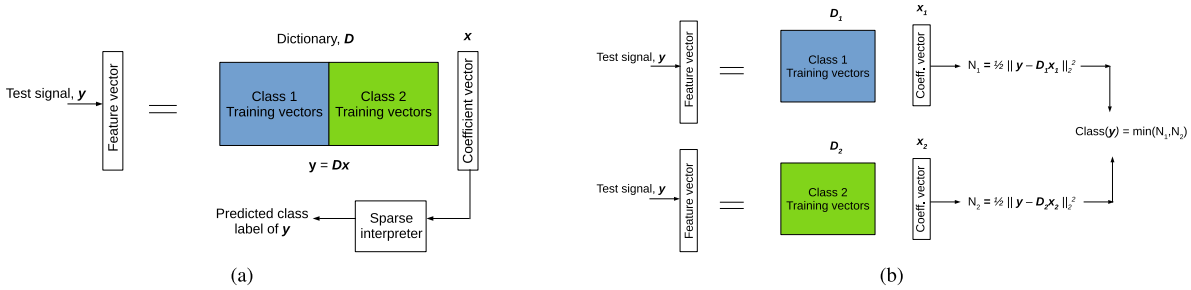

Fig. 1. Illustration of (a) existing and (b) proposed approach for binary classification. $\mathbf{D_i}$ is a dictionary consisting of training speech vectors from the ith class, N_i represent regression residual computed for the ith class. The predicted label of $\mathbf{y}$ corresponds to the class of dictionary which yields the minimum residual value.

where $\delta_i(\mathbf{x}) : \mathbb{R}_n \rightarrow \mathbb{R}_n$ returns the coefficients for class i. Its jth element is defined by

$$(\delta_i(\mathbf{x}))_j = \begin{cases} x_j & \text{if } \mathbf{d}_j \text{ in class } i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Finally, a class label p is assigned to $\mathbf{y}$, where

$$p = \underset{0 \leq i \leq C-1}{\operatorname{argmin}} \left(r_i(\mathbf{y})\right) \tag{6}$$

The SR classification approaches based on the l_1LS and NNLS sparse coding models will be shortly referred to as the l_1LS classification (LSRC) and NNLS classification (NSRC), respectively. While LSRC has been previously used in music genre classification [36] and speech recognition [35], the NNLS-based classifier has not been previously utilized in any speech-related tasks.

Both LSRC and NSRC use a single dictionary which consists of training exemplars from all the classes, as illustrated in Fig. 1(a). Compared to using exemplars stored in a single dictionary, using separate dictionaries for each class results in test exemplar being approximated as a linear combination of exemplars belonging to the same class only [38]. Also, in [37], the authors showed that solving a sparse coding problem separately by decomposing a large dictionary into multiple smaller dictionaries can reduce the computational burden. Therefore, in this work, we propose to organize the exemplars in separate dictionaries based on the class (healthy and PD) as shown in Fig. 1(b). Then, the test vectors are approximated as a non-negative linear combination of the exemplars in each of these dictionaries. Finally, the classification is performed by comparing the quality of reconstruction for different classes quantified by the regression residual and choosing the class which yields the minimum residual value as shown in Fig. 1(b). The LSRC and NSRC approach that utilize class-specific dictionaries are referred to in this study as Proposed-LSRC and Proposed-NSRC, respectively. Through experimental evaluation, we show that the use of separate class-specific dictionaries improves the overall PD detection performance.

III. Experimental Setup

A. Database Description

Two databases containing healthy and PD speech are used in this study. These databases are the widely used PC-GITA database [50], which represents a repository that has been recorded in a clinic under noise-controlled conditions using a professional microphone and audio card, and the MDVR-KCL database [51] that includes speech recorded by mobile devices. The details of these two databases are described below.

1) PC-GITA: This corpus contains speech signals collected using a variety of speech tasks by 50 PD patients (25 female and 25 male) and 50 control speakers (25 female and 25 male) whose native language is Colombian Spanish [50]. The PD patients have been diagnosed by neurologists. The healthy controls are free of any reported PD symptoms or other neuro-degenerative disease. The speaker age varies from 31 years old to 86 years old. The data was recorded in noise-controlled conditions in a sound proof booth using a dynamic omni-directional microphone (Shure, SM 63L) and sampled at 44.1 kHz with a resolution of 16 bits. The database includes speech collected using various speech tasks. The complete details of the database are available in [50]. In this study, we considered signals representing three speech tasks: (i) vowels (/a/, /e/, /i/, /o/, /u/), (ii) reading sentences aloud (six simple and complex sentences), and (iii) diadochokinetic (DDK) exercises (repetition of the sequence of syllables: /pa-ta-ka/, /pe-ta-ka/, /pa-ka-ta/, /pa/, /ka/, /ta/) from the database. Each speaker had uttered each vowel three times and uttered each DDK syllable and sentence one time. The speech signals were down-sampled to 16 kHz in order to be used in the experiments of the present study.

2) MDVR-KCL: The database consists of speech recordings from both early and advanced PD patients and HCs [51]. The corpus was collected using the Motorola Moto G4 Smartphone for two speech tasks (text reading and spontaneous dialogue) by 16 PD patients and 21 control speakers (except that the database available at [51] lacks the spontaneous dialogue sample of one of the PD patients). To perform the voice recordings on the device, a "Toggle Recording App" was developed, which uses the same functionalities as the voice recording module used within the i-PROGNOSIS Smartphone application, but works as a standalone Android application [51]. The voice capturing service runs as a standalone background service on the recording device and triggers voice recordings via on- and off-hook signals of the smartphone. The microphone signal is directly recorded (i.e. speech is not transmitted and therefore no low-bit-rate speech compression takes place) and high-quality recordings with a sample rate of 44.1 kHz and a bit depth of 16 bits are obtained [51]. The raw, uncompressed data is directly written

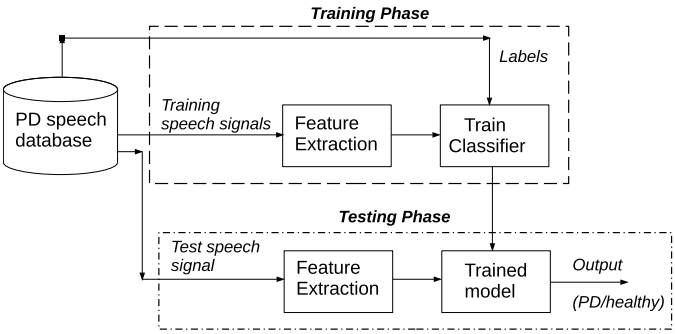

Fig. 2. The PD detection system based on the traditional machine learning approach consisting of two stages (the training phase and the testing phase).

to the external storage of the smartphone (SD-card) using the well-known WAVE file format (.wav). The speech signals were down-sampled to 16 kHz in order to be used in the experiments of the present study. For each speaker there is only one recording available for each speech task in the database. The length of the recording varies from approx. 1.2 min to 3.6 min. Most of the files contain long unnecessary pauses, unwanted sounds such as the mobile phone ringing and investigator's speech. These sections were manually removed with the help of the Audacity audio editor software. Furthermore, the cleaned signals were chopped into 3 s non-overlapping chunks to increase the size of the database. Altogether, for the text reading task, we obtained 534 speech segments from 21 healthy speakers and 423 speech segments from 16 PD patients. Similarly, we obtained 439 and 332 speech segments from 21 healthy and 15 PD patients for the spontaneous dialogue task, respectively.

B. Baseline Traditional ML-Based PD Detection Systems

In this work, we compare the performance of the SR classification approaches with traditional pipeline systems based on ML methods. The steps in developing a PD detection system based on the traditional ML pipeline approach are shown in Fig. 2. The system consists of two main parts: feature extraction and classifier.

1) Feature Extraction: In the feature extraction stage, selected features are extracted from the input speech signals. A wide variety of features have been proposed in the literature on PD detection [16], [17], [18], [19], [20], [21], [22]. In this study, we considered the INTERSPEECH 2010 paralinguistic challenge (IS10) feature set [43], which consists of various low-level descriptors (LLDs) characterizing three aspects of speech, namely, articulation, phonation, and prosody. The main reason for using features characterizing these three issues is that they are widely used in PD detection tasks [6], [23], [25], [26], and are also used as baselines in comparing various PD detection systems [6], [23], [26]. The feature set consists of 34 LLDs (such as MFCC, line spectral pair frequencies, fundamental frequency envelope) with 34 corresponding delta coefficients appended. 21 statistical functionals are applied to each of these 68 LLD contours. In addition, 19 functionals are applied to the 4 pitch-based LLD (such as jitter local, shimmer local) and their four delta coefficient contours. Finally the number of pitch onsets (pseudo syllables) and the total duration of the input are appended. In total, 1582 features are extracted from each speech file using the openSMILE toolkit [39]. The IS10 feature set consists of a combination of commonly used articulation, phonation and prosody features in speech-based PD detection studies. Hence these features are considered as "baseline features" in this study.

A few recent studies have analyzed the effectiveness of glottal source signals in PD detection [17], [24], [26]. These studies show that the glottal source waveform carries complementary PD-related information, and therefore combining the features derived from the glottal source with other features (articulation, phonation and prosody) can improve the PD detection performance. In this work, we propose to combine the baseline IS10 features with the MFCCs derived from the glottal source waveform to further enhance the performance of the PD detection systems. The computation of MFCCs from the glottal source is similar to that of the MFCC computation from speech, except that the input is the estimated glottal source waveform instead of the microphone speech signal. The glottal source waveforms were estimated using the quasi-closed phase (QCP) glottal inverse filtering method [44], [45]. 13-dimensional MFCCs (including the 0-th coefficient) were computed using 30 ms Hamming-windowed frames with a 5 ms shift. The MFCCs were computed from all the voiced frames of the glottal source signal form the MFCC parameter vector of the utterance. Four statistical measures were computed from the MFCC parameter vector: mean, standard deviation, kurtosis, and skewness. This results in $13 \times 4 = 52$ parameters representing the glottal MFCC feature set (gMFS). The glottal MFCCs can effectively capture glottal source variations in pathological speech signals.

2) Classifier: After feature extraction, an ML classifier is trained to distinguish PD speech from healthy speech. As mentioned earlier, several ML approaches have been studied for PD detection. For the purpose of this study, we chose two popular classifiers (SVM and RF), which are briefly described below.

- **SVM** [55] is the benchmark classifier employed in the automatic detection of diseases such as PD. The popularity of SVM is explained by the fact that speech databases recorded from patients usually contain little data, and SVM is very effective in classification tasks with limited training data [21], [47], [49]. In this work, we employed a non-linear SVM algorithm with a radial basis function (RBF) kernel. The kernel equation is given by

$$K(x, y) = exp(-\gamma||x - y||^2), \quad \gamma > 0, \tag{7}$$

  where x and y are training samples and labels, respectively, and γ is the kernel parameter. In addition to γ, regularization was used in the SVM with a regularization parameter C.
- **RF** [56] A random forest is an ensemble learning method that fits a number of decision tree classifiers to various sub-samples of the dataset and uses averaging to control over-fitting and provide stable predictions. The RF classifier contains a set of decision trees from a randomly selected subset of the training set. During testing, the votes from different decision trees are aggregated to determine

the final class of the test object. RF is one of the most used algorithms, because of its simplicity and diversity.

In the testing stage, as shown in Fig. 2, the trained ML models can be used to detect the presence of PD from speech signals. The same set of speech features that were used during training are extracted from test speech utterances. The extracted features are given as input to the classifiers, and the classifier predicts the labels ($healthy$ vs. PD). The baseline systems (shown in Fig. 2) developed with the RF and SVM classifiers are simply referred to as RF and SVM, respectively.

C. Evaluation Metrics

In this work, we consider five metrics, namely, recall, precision, F1-score, accuracy, and Matthews Correlation Coefficient metric (MCC), to evaluate the classification models. Recall is the ratio of the true positives to all (actual) positives in the data. Precision is the ratio of the correctly predicted positive examples divided by the total number of positive examples that were predicted. The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. A high F1-score represents a model that classifies well each observation into the correct class, indicating that the model will perform well on both precision and recall. Accuracy is the ratio of correct predictions to all predictions. The MCC metric is a contingency matrix method of calculating the Pearson product-moment correlation coefficient [53] between actual and predicted values. The range of values of MCC lies between -1 to +1. An MCC score of +1 and -1 indicates a perfect model and a poor model, respectively. A detailed description of considered metrics is available in [53], [54].

IV. Results

PD detection experiments were carried out using the proposed SR approaches (Proposed-NSRC and Proposed-LSRC), baseline SR approaches (LSRC and NSRC) and baseline ML systems (SVM and RF) with the IS10 feature set and the combined feature set (IS10+gMFS) described in Section III-B1. Distinct sets of SR, SVM, and RF classifiers were developed for each speech tasks of both the PC-GITA and MDVR-KCL databases. A 5-fold cross-validation (CV) strategy was followed, i.e., the speech data of 80% speakers was used for training and the data of the remaining 20% speakers was used for testing. There was no overlap of speakers used in training and testing, which guarantees speaker independence in the evaluation. The training data were z-score-normalized and the testing data were normalized by subtracting the mean and dividing by the standard deviation of the training sets for each feature. The optimal parameter values for the ML classifiers were derived using the Bayesian hyper-parameter optimization algorithm [52]. The optimization was performed following a 10-fold cross-validation strategy using the training data from the first fold. For SVM, the radial basis function kernel was used and the optimal values of box constraint and kernel scale were obtained using the optimization algorithm. In the case of RF, the optimization algorithm finds the optimal values for the three most important hyper-parameters: the number of trees in the forest, the maximum number of levels allowed in each tree, and the minimum number of samples required to be at a leaf node. Bayesian optimization takes an intelligent guess about the next combination to be tried by looking at the results of previous combinations. Whichever set of hyper-parameter produced better results, it will move towards those values. The best hyper-parameters obtained (based on accuracy) for each classifier following the optimization procedure were selected to be used for subsequent folds. The evaluation metrics were averaged over the 5 folds for evaluation. Note that the training data in each fold was exclusively used for training the ML classifiers. On the other hand, the data was simply pooled in the dictionary as column vectors for the SR classifiers, as they do not have separate training and testing steps.

A. Comparison of PD Detection Performances

Tables I–V show the results obtained for the PC-GITA and MDVR-KCL databases[1]. Table I shows the results obtained for the vowels of the PC-GITA database with the individual feature set (IS10) and combined feature set (IS10 + gMFS). From the table, it can be observed that in the case of the IS10 feature set, Proposed-NSRC provided the best performance in terms of recall (77.07%) accuracy (73.52%) and MCC (0.45). The SVM provided the best performance in terms of precision (72.19%). In terms of accuracy and MCC, the next best performing system was SVM. LSRC provided the lowest detection performance in terms of recall, precision, F1-score, accuracy and MCC. The same can be observed even with the combined feature set. The SR approaches based on NNLS (NSRC and Proposed-NSRC) performed better than their l_1LS counterparts (LSRC and Proposed-LSRC). With the combined feature set (IS10+gMFS), it can be clearly seen that there exists an improvement in performance for all the systems. This indicates the presence of complementary information among these two feature sets. Overall, the best performance in terms of precision (77.08%), F1-score (75.51%), accuracy (76%) and MCC (0.52) was obtained with Proposed-NSRC developed using the combined feature set.

Tables II and III show the results obtained for the DDK and sentence reading tasks of the PC-GITA database with the IS10 and the combined feature sets. With the DDK task, Proposed-NSRC gave the overall highest recall of 86.67%, F1-score of 83.20%, accuracy of 82.50% and MCC of 0.63, with the combined feature set. SVM was the second best performing system in terms of accuracy, F1-score and MCC. In case of the sentence reading task, Proposed-NSRC provided the overall best performance in terms of precision (84.51%), F1-score (83.17%), accuracy (82.84%) and MCC (0.64). As in the case of vowels, the combination of features further enhanced the performance of the individual detection systems. The performance of NSRC was close to that of the RF and SVM. For PC-GITA, it should be noted that all the systems provided better results for continuous speech (the DDK and sentence reading tasks) compared to vowels.

[1]The results obtained with glottal source features alone are not reported in the tables as these were inferior to those obtained with the openSMILE features alone or with the combination of the openSMILE and glottal features. Similar results have also been observed in previous studies [26], [47], [49].

TABLE I
RESULTS FOR THE VOWELS IN THE PC-GITA DATABASE WITH THE INDIVIDUAL CLASSIFIERS AND FEATURE SETS

Classifier	IS10					IS10+gMFS				
	Recall	Precision	F1-score	Accuracy	MCC	Recall	Precision	F1-score	Accuracy	MCC
ML classification approaches										
RF	72.00	70.13	71.05	70.67 ± 3.92	0.42 ± 0.08	70.40	74.13	72.80	72.01 ± 3.98	0.46 ± 0.10
SVM	72.67	**72.19**	72.43	72.33 ± 3.62	0.45 ± 0.08	73.33	75.74	74.87	74.34 ± 3.96	0.50 ± 0.08
SR classification approaches										
LSRC	64.00	63.99	63.93	63.72 ± 3.84	0.28 ± 0.10	64.53	64.10	64.27	64.06 ± 4.03	0.29 ± 0.15
NSRC	71.87	66.87	68.20	69.25 ± 3.79	0.37 ± 0.11	**77.47**	70.83	72.80	73.93 ± 4.05	0.46 ± 0.12
Proposed-LSRC	73.47	67.77	69.27	70.49 ± 3.81	0.39 ± 0.10	73.20	68.31	69.60	70.65 ± 4.22	0.39 ± 0.14
Proposed-NSRC	**77.07**	70.42	**73.59**	**73.52 ± 3.79**	**0.45 ± 0.09**	74.00	**77.08**	**75.51**	**76.00 ± 4.09**	**0.52 ± 0.10**

Except for *MCC* all metrics are in (%).

TABLE II
RESULTS FOR THE DDK UTTERANCES IN THE PC-GITA DATABASE WITH THE INDIVIDUAL CLASSIFIERS AND FEATURE SETS

Classifier	IS10					IS10+gMFS				
	Recall	Precision	F1-score	Accuracy	MCC	Recall	Precision	F1-score	Accuracy	MCC
ML classification approaches										
RF	77.67	76.87	76.67	77.01 ± 4.27	0.54 ± 0.08	75.33	76.92	76.00	75.95 ± 4.22	0.52 ± 0.07
SVM	**85.00**	76.12	80.31	79.17 ± 4.14	0.59 ± 0.10	78.33	**82.46**	80.34	80.83 ± 4.18	0.61 ± 0.06
SR classification approaches										
LSRC	80.33	70.14	72.50	74.45 ± 4.58	0.46 ± 0.07	81.33	69.94	72.67	74.78 ± 4.35	0.47 ± 0.06
NSRC	83.33	75.76	79.37	78.33 ± 4.32	0.56 ± 0.08	87.00	75.47	79.17	80.77 ± 4.20	0.59 ± 0.08
Proposed-LSRC	78.00	75.31	76.00	76.55 ± 4.32	0.52 ± 0.06	79.67	75.55	76.83	77.47 ± 4.22	0.54 ± 0.05
Proposed-NSRC	83.33	**78.13**	**80.65**	**80.00 ± 4.19**	**0.60 ± 0.10**	**86.67**	80.00	**83.20**	**82.50 ± 4.23**	**0.63 ± 0.06**

Except for *MCC* all metrics are in (%).

TABLE III
RESULTS FOR THE SENTENCES IN THE PC-GITA DATABASE WITH THE INDIVIDUAL CLASSIFIERS AND FEATURE SETS

Classifier	IS10					IS10+gMFS				
	Recall	Precision	F1-score	Accuracy	MCC	Recall	Precision	F1-score	Accuracy	MCC
ML classification approaches										
RF	81.67	76.56	79.03	78.33 ± 4.12	0.59 ± 0.06	80.33	81.62	80.83	80.80 ± 3.84	0.62 ± 0.08
SVM	83.33	78.13	80.65	80.00 ± 4.18	0.61 ± 0.09	83.33	79.76	80.83	81.41 ±3.79	0.62 ± 0.10
SR classification approaches										
LSRC	81.00	68.48	71.50	74.00 ± 4.32	0.44 ± 0.10	80.67	69.07	71.83	74.20 ± 3.96	0.45 ± 0.08
NSRC	**85.00**	76.12	80.31	79.17 ± 4.14	0.58 ± 0.11	**83.33**	78.13	80.65	80.00 ± 3.84	0.60 ± 0.09
Proposed-LSRC	80.33	74.39	76.17	77.12 ± 4.27	0.53 ± 0.05	81.33	75.20	77.00	77.99 ± 3.92	0.54 ± 0.06
Proposed-NSRC	80.00	**82.76**	**81.36**	**81.67 ± 4.22**	**0.63 ± 0.10**	81.67	**84.51**	**83.17**	**82.84 ± 3.81**	**0.64 ± 0.08**

Except for *MCC* all metrics are in (%).

TABLE IV
RESULTS FOR THE UTTERANCES FROM THE TEXT READING TASK IN THE MDVR-KCL DATABASE WITH THE INDIVIDUAL CLASSIFIERS AND FEATURE SETS

Classifier	IS10					IS10+gMFS				
	Recall	Precision	F1-score	Accuracy	MCC	Recall	Precision	F1-score	Accuracy	MCC
ML classification approaches										
RF	**88.74**	77.72	82.81	77.48 ± 7.21	0.46 ± 0.10	88.70	75.72	81.69	81.19 ± 6.84	0.48 ± 0.08
SVM	84.37	75.49	79.68	77.78 ± 8.32	0.45 ± 0.08	88.84	79.20	83.50	78.48 ± 7.06	0.46 ± 0.07
SR classification approaches										
LSRC	78.03	74.64	76.30	69.23 ± 7.43	0.32 ± 0.10	78.79	77.04	77.90	71.63 ± 6.92	0.38 ± 0.05
NSRC	88.16	77.26	82.02	76.31 ± 8.09	0.42 ± 0.08	82.31	**88.97**	85.51	78.19 ± 6.97	0.42 ± 0.07
Proposed-LSRC	71.00	80.43	75.42	75.15 ± 6.97	0.41 ± 0.07	77.26	88.16	82.02	76.31 ± 6.79	0.42 ± 0.06
Proposed-NSRC	88.41	**80.60**	**83.71**	**78.88 ± 7.33**	**0.47 ± 0.08**	**89.84**	82.73	**86.14**	**82.46 ± 7.06**	**0.50 + 0.09**

Except for *MCC* all metrics are in (%).

TABLE V
RESULTS FOR THE UTTERANCES FROM THE SPONTANEOUS DIALOGUE TASK IN THE MDVR-KCL DATABASE WITH THE INDIVIDUAL CLASSIFIERS AND FEATURE SETS

Classifier	IS10					IS10+gMFS				
	Recall	Precision	F1-score	Accuracy	MCC	Recall	Precision	F1-score	Accuracy	MCC
ML classification approaches										
RF	83.33	76.53	79.79	78.29 ± 4.57	0.53 ± 0.08	83.33	76.92	79.17	80.19 ± 4.63	0.54 ± 0.07
SVM	83.92	72.52	75.04	76.79 ± 4.73	0.51 ± 0.09	83.33	76.53	79.79	78.29 ± 4.89	0.53 ± 0.08
SR classification approaches										
LSRC	82.66	69.93	73.50	70.02 ± 4.82	0.42 ± 0.05	**93.33**	67.20	78.14	73.14 ± 4.73	0.49 ± 0.06
NSRC	**87.06**	76.29	81.32	77.56 ± 4.80	0.53 ± 0.06	88.54	75.09	75.41	80.17 ± 4.80	0.53 ± 0.08
Proposed-LSRC	83.76	75.07	76.85	74.70 ± 4.73	0.52 ± 0.06	85.51	65.56	74.21	76.57 ± 4.54	0.53 ± 0.10
Proposed-NSRC	84.83	**81.00**	**82.63**	**81.94 ± 4.63**	**0.55 ± 0.07**	82.46	**79.66**	**81.03**	**83.08 ± 4.58**	**0.57 ± 0.09**

Except for *MCC* all metrics are in (%).

Compared to vowels, continuous speech contains richer dynamic information about prosody, articulation of various phonemes, and transitions between different linguistic units, which helps in better characterization of PD. The DDK and sentence tasks achieved a similar performance in the automatic detection of PD. It should be noted, however, that the DDK tasks followed in the collection of the syllable-rhythmic Spanish PC-GITA dataset include continuous repetitions of syllable sequences (/pa-ta-ka/, /pe-ta-ka/, /pa-ka-ta/, /pa/, /ka/, /ta/). For languages, which are not strongly syllable-rhythmic, certain DDK tasks may not be well suited to differentiate PD from HC [26]. Therefore, it might be that the current results related to the DDK task cannot be fully generalised to other languages.

In the literature, several PD detection studies have reported results based on the PC-GITA database. However, the classification results (e.g. accuracy) are not consistent between individual investigations. This is due to differences in, for example, the considered speech tasks and experimental setups. For the vowels of PC-GITA, a few recent studies reported accuracies around 90% [60], [61], [62]. However, these studies considered each vowel of PC-GITA separately in training and testing the classification models. For example, in [60], the highest accuracy of 91% was achieved with the model which was trained and tested using only the vowel /a/. It should be noted that in the current study all the vowels were considered together in building a classification model. Therefore, a direct comparison of the present results with those reported, for example, in [60] is not justified as it can potentially mislead the reader to understand that the results of the present study are worse than in previously reported similar studies. In a recent study [21], SVM classifiers were trained by considering all the vowels of PC-GITA together as in the current study, and the authors reported the best accuracy of 76% using SFF-based features. The best accuracy (76%) obtained in this study for the vowel production task with the proposed-NSRC approach is the same as the one reported in [21]. Furthermore, the best accuracy (82.5%) obtained with the proposed-NSRC approach for the DDK task is better than the accuracy of 76% reported in [29]. Note that the authors of [21] have not evaluated their models using the DDK task, and the authors of [29] have not evaluated their models using the vowel production task. Therefore, in comparison with [21], [29], the proposed-NSRC approach provides comparable or better performance in discriminating healthy speakers from PD patients.

Furthermore, we conducted post analysis of the accuracies received for all CV folds with the proposed-NSRC approach in case of DDK and sentence tasks, in order to preliminarily assess the clinical utility of the proposed-NSRC approach as an early PD screening tool. In the PC-GITA database, the speech signals of each PD patient are assigned with an Unified Parkinson's Disease Rating Scale (UPDRS) speech score varying between 0 and 3 according to the level of speech impairment. For post analysis, we grouped the PD patients into three classes: 0-patients with mildly affected speech (UPDRS speech score $= 0$), 1-patients with moderately affected speech (UPDRS speech score $= 1$), and 2-patients with severe speech deficits (UPDRS speech score $>$ 1). The analysis results indicated that the PD detection accuracy varied between 67%–74% for mild cases, between 83%–88% for moderate cases and between 85%–92% for severe cases. This indicates that the proposed approach has better identification accuracy for moderate and severe PD cases compared to mild case. In mild cases, the speech of PD patients is only slightly impaired which posses difficulty in distinguishing PD patients from healthy individuals. We argue, however, that the obtained identification accuracies for mild case can be considered good enough for the potential use of the proposed-NSRC approach even in early detection of PD.

Tables IV and V show the results obtained for the text reading and spontaneous speech tasks of the MDVR-KCL database. Note that the MDVR-KCL database is moderately imbalanced in terms of the number of speakers and utterances for each class. Furthermore, the demographic information is unfortunately not available for the MDVR-KCL database. Hence, we would like to remind the reader that the possibility that gender or age might have influenced the results obtained with this database cannot be completely ruled out. Results similar to PC-GITA are also observed for the MDVR-KCL database. From Table IV, it can be observed that in the case of the IS10 feature set, Proposed-NSRC provided the best performance in terms of precision (81.00%), F1-score (83.71%), MCC (0.47) and accuracy (78.88%). RF provided the next best performance in terms of precision (77.720%), F1-score (82.81%), MCC (0.46%), accuracy (77.48%). The recall values provided by RF and Proposed-NSRC are close to each other. With the combined feature set, Proposed-NSRC provided the overall best performance yielding a recall of 89.84%, F1-score of 86.14%, accuracy of 82.46% and MCC of 0.53. From Table V, it can be seen that as in the results obtained for the text reading task, Proposed-NSRC provided the best overall result with an F1-score of 81.03%, precision of 79.66%, accuracy of 83.08% and MCC of 0.57 for the spontaneous speech task. It can be seen that the performance of the detection systems improves when the glottal features are combined with the IS10 feature set due to the complementary information among the feature sets. An important observation is that none of the considered classification approaches were able to show the best recall or precision value consistently for all speech tasks, feature sets and databases. If a model of either high recall or high precision is to be used, for example, as a PD screening tool in clinical diagnosis, it is not straightforward to chose the best one among the considered classification frameworks. However, we argue that Proposed-NSRC is a better choice as a screening tool since it consistently provided either the best recall or precision, and also strikes the best balance between precision and recall as indicated by its higher F1-scores for all cases.

It is worth noting that SVM performed better than RF in terms of F1-score and accuracy for the balanced PC-GITA database, but RF provided better performance than SVM for the imbalanced MDVR-KCL database. However, Proposed-NSRC achieved better performance than RF and SVM for both of the databases in terms of F1-score and accuracy. For imbalanced/balanced datasets, the MCC metric is considered more reliable than accuracy and F1-score [54] because MCC produces a high score only if the prediction shows good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally to the size of both classes in the dataset [54]. From Tables I, II, III, IV, and V, it can be seen that Proposed-NSRC achieved

better MCC values than the other systems. This indicates the effectiveness of Proposed-NSRC for balanced as well as moderately imbalanced datasets.

The common trends in the results obtained for both databases can be summarized as follows. First, for all speech tasks, there is an improvement in the performance for all the detection systems when the combined feature set is used. This is because the MFFCs computed from the glottal source signal carry effective information about phonation. Hence, combining the glottal MFCCs with IS10 (which contain information about articulation, prosody and phonation) can better characterize speech of PD patients. Second, the SR approaches using class-specific dictionaries (Proposed-LSRC and Proposed-NSRC) gave better detection performance than the SR approaches that use a single dictionary (LSRC and NSRC). Third, the NNLS approaches (NSRC and Proposed-NSRC) achieved much better detection performance compared to the l_1LS approaches (LSRC and Proposed-LSRC). Overall, the PD detection performance achieved with Proposed-NSRC is better than that of all other compared systems, for both of the considered databases. Moreover, the best PD detection performance for the PC-GITA database was achieved (F1-score of 83.17%, accuracy of 82.84% and MCC of 0.64) when the combined feature set extracted from the sentence utterances was used with Proposed-NSRC. Similarly, in case of the MDVR-KCL database, the highest values (accuracy of 83.08% and MCC of 0.57) were given by Proposed-NSRC which was developed using the combined feature set computed from the utterances of the spontaneous dialogue task.

V. Conclusion

People with PD commonly suffer from speech disorders and communication problems. Detection of PD at an early stage of the disease is essential and can be approached using automatic speech-based classification. In this work, we investigated the use of the SR classification approaches based on the NNLS and l_1LS sparse coding models to identify people with PD using speech signals. In the SR approach, the sparse vector obtained for the test feature vector is interpreted to predict the class label. Unlike in the existing SR approaches (LSRC and NSRC), in this work, we organized the exemplars in separate dictionaries based on the class (the associated health status) and used them to approximate test exemplar as a linear combination of the exemplars in each of these dictionaries. The exemplars were the feature sets (IS10/IS10+gMFS) derived from speech signals. Finally, classification (healthy vs. PD) was performed by finding the class sequence yielding the minimum reconstruction error between the test exemplar and its estimate.

The experiments were conducted using two databases (PC-GITA and MDVR-KCL) and using two features sets (IS10 and IS10+gMFS). It was observed that the NNLS approaches were better than the l_1LS approaches in the detection of PD. The results showed that using class-specific dictionaries results in improvement of detection performance, and that Proposed-NSRC outperformed the other SR approaches. Furthermore, Proposed-NSRC delivered consistently better overall performance compared to the baseline systems in discriminating PD speech from healthy speech for both databases. The overall results suggest that the usage of Proposed-NSRC can be considered beneficial and promising as it avoids the tedious training phase and hyper-parameter tuning, but still better discriminates healthy and PD speech from both clean as well as mobile device recordings. The current study can be viewed as a reference point to carry out further research in PD detection, or in general for voice pathology detection, using SR techniques, particular those based on the NNLS sparse coding model. In the future, SR-based techniques can be used to predict the dysarthria level of patients with PD. The Proposed-NSRC approach in its current form is not suitable for regression task. Hence, future investigations may focus on how it can be adopted for regression problems such as predicting the unified Parkinson disease rating scale (UPDRS) to monitor the disease progression. Furthermore, the performance of SR-based techniques can be investigated in the detection of other speech disorders such as dysphonia and can be used in multi-class classification.

References

[1] J. Parkinson, "An essay on the shaking palsy," *J. Neuropsychiatry Clin. Neurosci.*, vol. 14, no. 2, pp. 223–236, 2002.

[2] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behav. Neurol.*, vol. 11, no. 3, pp. 131–137, 1998.

[3] J. S. Almeida et al., "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognit. Lett.*, vol. 125, pp. 55–62, 2019.

[4] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain Cogn.*, vol. 56, no. 1, pp. 24–29, 2004.

[5] R. E. Burke, "Evaluation of the braak staging scheme for Parkinson's disease: Introduction to a panel presentation," *Movement Disord.*, vol. 25, no. S1, pp. S76–S77, 2010.

[6] 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, 2021, Art. no. 102418.

[7] Y. Camnos-Roca, F. Calle-Alonso, C. J. Perez, and L. Naranjo, "Computational diagnosis of Parkinson's disease from speech based on regularization methods," in *Proc. IEEE 26th Eur. Signal Process. Conf.*, 2018, pp. 1127–1131.

[8] J. Vasquez et al., "Convolutional neural networks and a transfer learning strategy to classify Parkinson's disease from speech in three different languages," in *Proc. 24th Iberoamerican Congr. Pattern Recognit. Image Anal., Comput. Vis., Appl.*, 2019, pp. 697–706.

[9] C. O. Sakar et al., "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Appl. Soft Comput.*, vol. 74, pp. 255–263, 2019.

[10] A. Benba, J. Abdelilah, and A. Hammouch, "Voice analysis for detecting persons with Parkinson's disease using PLP and VQ," *J. Theor. Appl. Inf. Technol.*, vol. 70, no. 3, pp. 443–450, 2014.

[11] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect Parkinson's disease from speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 1155–1159.

[12] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[13] B. E. Sakar et al., "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE J. Biomed. Health Inform.*, vol. 17, no. 4, pp. 828–834, Jul. 2013.

[14] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[15] M. Shahbakhi, D. T. Far, and E. Tahami, "Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine," *J. Biomed. Sci. Eng.*, vol. 2014, pp. 147–156, 2014.

[16] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička, "Automatic evaluation of articulatory disorders in Parkinson's disease," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014.

[17] 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. Methods Programs Biomed.*, vol. 154, pp. 89–97, 2018.

[18] Z. Galaz et al., "Changes in phonation and their relations with progress of Parkinson's disease," *Appl. Sci.*, vol. 8, no. 12, 2018, Art. no. 2339.

[19] Z. Cai et al., "An intelligent Parkinson's disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy k-NN approach," *Comput. Math. Methods Med.*, vol. 2018, 2018, Art. no. 2396952.

[20] J. R. Orozco-Arroyave et al., "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *J. Acoust. Soc. Amer.*, vol. 139, no. 1, pp. 481–500, Jan. 2016.

[21] S. R. Kadiri, R. Kethireddy, and P. Alku, "Parkinson's disease detection from speech using single frequency filtering cepstral coefficients," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 4971–4975.

[22] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 10, pp. 1100–1108, Oct. 2016.

[23] F. O. López-Pabón, T. Arias-Vergara, and J. R. Orozco-Arroyave, "Cepstral analysis and Hilbert-Huang transform for automatic detection of Parkinson's disease," *TecnoLógicas*, vol. 23, no. 47, pp. 91–106, 2020.

[24] M. Novotný, P. Dusek, I. Daly, E. Ruzicka, and J. Rusz, "Glottal source analysis of voice deficits in newly diagnosed drug-naive patients with Parkinson's disease: Correlation between acoustic speech characteristics and non-speech motor performance," *Biomed. Signal Process. Control*, vol. 57, no. 1, pp. 1–9, 2020.

[25] T. Arias-Vergara, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, P. Klumpp, and E. Nöth, "Unobtrusive monitoring of speech impairments of Parkinson's disease patients through mobile devices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6004–6008.

[26] N. P. Narendra, B. Schuller, and P. Alku, "The detection of Parkinson's disease from speech using voice source information," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1925–1936, 2021.

[27] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with Parkinson's disease," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 314–318.

[28] J. R. Orozco-Arroyave, Analysis of speech of people with Parkinson's disease (doctoral thesis), Ph.D. dissertation, Logos Verlag Berlin GmbH, Berlin, Germany, 2016.

[29] J. C. Vásquez-Correa, C. D. Rios-Urrego, A. Rueda, J. R. Orozco-Arroyave, S. Krishnan, and E. Nöth, "Articulation and empirical mode decomposition features in diadochokinetic exercises for the speech assessment of Parkinson's disease patients," in *Proc. 24th Iberoamerican Congr., Prog. Pattern Recognit., Image Anal., Comput. Vis., Appl.*, 2019, pp. 688–696.

[30] 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.

[31] X. Hang and F.-X. Wu, "Sparse representation for classification of tumors using gene expression data," *J. Biomed. Biotechnol.*, vol. 2009, 2009, Art. no. 403689.

[32] Y. Li and A. Ngom, "Classification approach based on non-negative least squares," *Neurocomputing*, vol. 118, pp. 41–57, 2013.

[33] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[34] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[35] D. Baby, T. Virtanen, J. F. Gemmeke, and H. Van hamme, "Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1788–1799, Nov. 2015.

[36] X. Chen and P. J. Ramadge, "Music genre classification using multiscale scattering and sparse representations," in *Proc. IEEE 47th Annu. Conf. Inf. Sci. Syst.*, 2013, pp. 1–6.

[37] T. Ge, K. He, and J. Sun, "Product sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 939–946.

[38] E. Yılmaz and J. F. Gemmeke, "Noise robust exemplar matching using sparse representations of speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, 1306–1319, Aug. 2014.

[39] "openSMILE 3.0.1," 2022. [Online]. Available: https://github.com/audeering/opensmile/releases

[40] Y. Li and A. Ngom, "Sparse representation approaches for the classification of high-dimensional biological data," *BMC Syst. Biol.*, vol. 7, no. 4, pp. 1–14, 2013.

[41] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2006, pp. 609–616.

[42] Y. Li, "Sparse representation toolbox in MATLAB," 2015. [Online]. Available at: https://sites.google.com/site/sparsereptool/

[43] B. Schuller et al., "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2794–2797.

[44] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, Jul. 1991.

[45] P. Alku et al., "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, May 2002.

[46] M. Airas et al., "A toolkit for voice inverse filtering and parametrisation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 2145–2148.

[47] M. K. Reddy et al., "The automatic detection of heart failure using speech signals," *Comput. Speech Lang.*, vol. 69, 2021, Art. no. 101205.

[48] M. K. Reddy et al., "Glottal flow characteristics in vowels produced by speakers with heart failure," *Speech Commun.*, vol. 137, pp. 35–43, 2022.

[49] M. K. Reddy, P. Alku, and K. S. Rao, "Detection of specific language impairment in children using glottal source features," *IEEE Access*, vol. 8, pp. 15273–15279, 2020.

[50] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2014, pp. 342–347.

[51] H. Jaeger et al., "Mobile device voice recordings at King's college london (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls," 2019. Accessed: May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.2867216

[52] D. J. Lizotte, "Practical Bayesian optimization," Ph.D. dissertation, Univ. of Alberta, Edmonton, AB, Canada, 2008.

[53] S. Boughorbel et al., "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS One*, vol. 12, no. 6, 2017, Art. no. e0177678.

[54] S. Y. Wong et al., "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.

[55] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[56] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[57] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[58] J. I. Godino-Llorente et al., "Towards the identification of idiopathic Parkinson's disease from the speech. new articulatory kinetic biomarkers," *PLos One*, vol. 12, 2017, Art. no. e0189583.

[59] L. Moro-Velazquez, G.-G. Jorge Andrés, G.-L. Juan Ignacio, V. Jesus, O.-A. Juan Rafael, and D. Najim, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease," *Appl. Soft Comput.*, vol. 62, pp. 649–666, 2018.

[60] B. Karan and S. Sekhar Sahu, "An improved framework for Parkinson's disease prediction using variational mode decomposition-Hilbert spectrum of speech signal," *Biocybernetics Biomed. Eng.*, vol. 41, no. 2, pp. 717–732, 2021.

[61] 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, 2020, Art. no. 102050.

[62] B. Karan et al., "Parkinson disease prediction using intrinsic mode function based features from speech signal," *Biocybernetics Biomed. Eng.*, vol. 40, no. 1, pp. 249–264, 2020.

Mittapalle Kiran Reddy received the M.E. degree in communication systems from the SSN College of Engineering, Chennai, India, in 2014, and the Ph.D. degree in speech processing from the Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Kharagpur, Kharagpur, India, in 2019. From October 2014 to March 2018, he was a Senior Scientific Officer in the research project sponsored by the Department of Information Technology, Goverment of India, undertaken by IIT Kharagpur. He is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include signal processing, speech synthesis, speech recognition, analysis and detection of speech disorders, and machine learning.

Paavo Alku (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc.(Tech). degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and an Assistant Professor and Professor with the University of Turku, Turku, Finland, from 1994 to 1999. He is currently a Professor of speech communication technology with Aalto University, Espoo, Finland. He has authored or coauthored around 230 peer-reviewed journal articles and around 220 peer-reviewed conference papers. He is an Associate Editor for *Journal of the Acoustical Society of America*. His research interests include analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modelling of speech, speech-based biomarking of human health, and cerebral processing of speech. He was an Academy Professor assigned by the Academy of Finland during 2015–2019. He is a Fellow of the ISCA.