Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings

Article *in* IEEE Journal of Biomedical and Health Informatics · July 2013

DOI: 10.1109/JBHI.2013.2245674

Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings

Betul Erdogdu Sakar, M. Erdem Isenkul, C. Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun

***Abstract*—There has been an increased interest in speech pattern analysis applications of Parkinsonism for building predictive telediagnosis and telemonitoring models. For this purpose, we have collected a wide variety of voice samples, including sustained vowels, words, and sentences compiled from a set of speaking exercises for people with Parkinson's disease. There are two main issues in learning from such a dataset that consists of multiple speech recordings per subject: 1) How predictive these various types, e.g., sustained vowels versus words, of voice samples are in Parkinson's disease (PD) diagnosis? 2) How well the central tendency and dispersion metrics serve as representatives of all sample recordings of a subject? In this paper, investigating our Parkinson dataset using well-known machine learning tools, as reported in the literature, sustained vowels are found to carry more PD-discriminative information. We have also found that rather than using each voice recording of each subject as an independent data sample, representing the samples of a subject with central tendency and dispersion metrics improves generalization of the predictive model.**

***Index Terms*—Central tendency and dispersion metrics, cross validation, multiple sound types, speech impairments, telediagnosis of Parkinson's disease.**

I. Introduction

PARKINSON's disease (PD) is a neurodegenerative disorder of central nervous system that causes partial or full loss in motor reflexes, speech, behavior, mental processing, and other vital functions [1]. In 1817, PD was described as "shaking palsy" by Doctor James Parkinson [2]. It is generally observed in elderly people and causes disorders in speech and motor abilities (writing, balance, etc.) of 90% of the patients [3]. Ensuing Alzheimer, PD is the second common neurological health problem in elder ages and it is estimated that nearly 10 million people all around the world and approximately 100 000 in Turkey are suffering from this disease [4], [5]. Particularly, PD is generally seen in one out of every hundred people aged over 65. Currently, there is no known cure for the disease [6], [7]. Although, there is significant amount of drug therapies to decrease difficulties caused by the disorder, PD is usually diagnosed and treated using invasive methods [8]. Therefore, this complicates the process of diagnosis and treatment of patients who are grieving from the disease.

IN this study, using speech data from subjects is expected to help the development of a noninvasive diagnostic. There are important examples of these kinds of Alzheimer and PD studies all around the world. The studies based on the PD focus on symptoms like slowness in movement, poor balance, trembling, or stiffness of some body parts [9]–[12] but especially voice problems. The main reason behind the popularity of PD diagnosis from speech impairments is that telediagnosis and telemonitoring systems based on speech signals are low in cost and easy to self-use [7], [13]. Such systems lower the inconvenience and cost of physical visits of PD patients to the medical clinic, enable the early diagnosis of the disease, and also lessen the workload of medical personnel [7], [13]–[15]. People with Parkinsonism (PWP) suffer from speech impairments like dysphonia (defective use of the voice), hypophonia (reduced volume), monotone (reduced pitch range), and dysarthia (difficulty with articulation of sounds or syllables). Even though there are many studies aiming at diagnosing and monitoring PD using these impairments, the origin of these studies leans to diagnose basic voice disorders.

Voice disorders can be measured by acoustic tools simply using aperiodic vibrations in the voice. Complex nonlinear aperiodicity, and turbulent, aeroacoustic, non-Gaussian randomness of the sound can be used to increase the clinical usefulness of the voice disorder diagnosis systems [13]. As for impairments, Little *et al*. [7] aimed to analyze the stage of the disease by measuring the dysphonia caused by PD. In their study, sustained vowel "a" phonations were recorded from 31 subjects of which 23 were diagnosed with PD. Then, dysphonia measures were extracted from the phonations to identify the grade of the disease by telemonitoring. Tsanas *et al*. [16] used speech data to predict tendency for progression of disease in which they extracted voice features using signal processing algorithms from 6000 samples of 42 PWP and identified the useful features. Selected

Manuscript received September 16, 2012; revised January 9, 2013; accepted January 29, 2013. Date of publication February 6, 2013; date of current version June 27, 2013. This work was supported in part by Scientific Research Projects Coordination Unit of Istanbul University under Grant 11002 and in part by Bogazici University Scientific Research Project under Grant 6392. The work of B.E. Sakar and C.O. Sakar was supported by the Ph.D. Scholarship (2211) from Turkish Scientific Technical Research Council (TÜBITAK).

B. E. Sakar is with the Department of Computer Programming, Bahcesehir University, Istanbul 34381, Turkey (e-mail: betul.erdogdu@bahcesehir.edu.tr).

M. E. Isenkul, A. Sertbas, and O. Kursun are with the Department of Computer Engineering, Istanbul University, Istanbul 34320, Turkey (e-mail: eisenkul@istanbul.edu.tr; asertbas@istanbul.edu.tr; okursun@istanbul.edu.tr).

C. O. Sakar is with the Department of Computer Engineering, Bahcesehir University, Istanbul 34353, Turkey (e-mail: okan.sakar@bahcesehir.edu.tr).

F. Gurgen is with the Department of Computer Engineering, Bogazici University, Istanbul 34342, Turkey (e-mail: gurgen@boun.edu.tr).

S. Delil and H. Apaydin are with the Department of Neurology, Cerrahpaşa Faculty of Medicine, Istanbul University, Fatih 34098, Turkey (e-mail: sakirdelil@yahoo.com; hulyapay@istanbul.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2013.2245674

subsets of features were mapped to UPDRS (Unified Parkinson's Disease Rating Scale) using regression and classification techniques. Another study that makes a selection of dysphonia measure subsets used support vector machines (SVM) to distinguish 263 samples belonging to 43 subjects [14]. Contrary to the traditional studies, nonlinear speech analysis algorithms can also be used with standard PD metrics like UPDRS or Hoehn & Yahr Scale (H&Y). In such a study [15], using nonlinear speech analysis algorithms with UPDRS scores hold on ∼6000 recordings from 42 PWP ended up with a slightly different estimation (about 2 points) from the clinicians' UPDRS estimates. UPDRS and H&Y are taken into consideration to study the statistical relationship between the two metrics and find an improvement in H&Y estimation [17]. Sakar and Kursun [18] applied a mutual information-based feature selection algorithm with the permutation test for assessing the relevance and the statistical significance of the relations between the features and the PD-score, fed the selected features into an SVM for building a classification model, and used a leave-one-subject-out (LOSO) cross-validation scheme to avoid bias. For classification, most studies use SVM to distinguish healthy subjects from PWP [7], [18], [19] and the success of the diagnosis system is measured with ROC curves, true positive and false positive rates [20].

The purpose of this study is to design a computer-aided data collection, storage, and analysis system to simplify the process of diagnosis and treatment of PD in the Department of Neurology in Cerrahpaşa Faculty of Medicine, Istanbul University. Speech recordings, demographic information, health background, and progression of PD of each patient is gathered and stored. Then, collected speech recordings are parsed and a series of features are extracted from the voice samples. The speech datasets used in the field of PD diagnosis, as well as this study, generally consist of multiple speech recordings per subject [21]. The dataset collected in this study contains multiple voice samples per subject such as sustained vowels, numbers, words, and short sentences. In this paper, we also compare the success of alternative cross-validation techniques that can be used with such datasets in a classification algorithm built for PD diagnosis. As classification algorithms, we use k-nearest neighbor (k-NN) and support vector machines (SVM) and evaluate the success of the models in discriminating the healthy subjects from subjects with PD according to their accuracy, specificity, sensitivity, and Matthews correlation coefficient (MCC) scores.

This paper is organized as follows: In Section II, the data acquisition techniques and the collected speech dataset are described. Section III summarizes the overall structure of the built PD diagnosis system. Section IV presents the experimental results. We present the conclusions and discussions in Section V.

II. Data Acquisition

The data collected in the context of this study (Fig. 1) belongs to 20 PWP (6 female, 14 male) and 20 healthy individuals (10 female, 10 male) who appealed at the Department of Neurology in Cerrahpaşa Faculty of Medicine, Istanbul University. Test group consists of patients who are suffering from PD for 0 to 6 years. Individual ages vary between 43 and 77 (mean: 64.86, standard deviation: 8.97) along with 45 and 83 (mean: 62.55, standard deviation: 10.79) for test and control groups, respectively. From all subjects, 26 voice samples including sustained vowels, numbers, words, and short sentences are recorded. The voice samples are selected by a group of neurologists from a set of speaking exercises that aim to lead to more powerful sound of PWP [22]. Recording is achieved by a Trust MC-1500 microphone with a frequency range between 50 Hz and 13 kHz. The microphone is set to 96 kHz, 30 dB and placed at 10 cm distant from the subject and then the subject is asked to read or repeat the specified text.

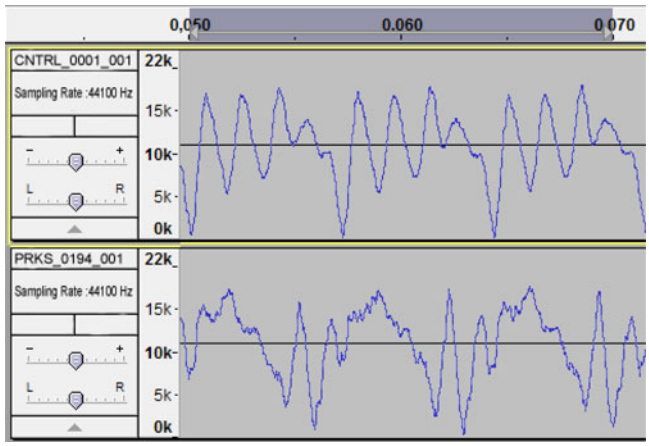

Fig. 1. Waveform of a voice sample belonging to a PWP (bottom) and a healthy individual (top). Amplitude of the signal (y-axis) is plotted against time duration (x-axis).

After collecting the aforementioned dataset with multiple types of sound recordings and performing our experiments, in line with the obtained findings, we continued collecting an independent test set from PWP via the same physician's examination process under the same conditions. During the collection of this dataset, 28 PD patients are asked to say only the sustained vowels "a" and "o" three times, respectively which makes a total of 168 recordings. Test group consists of patients who are suffering from PD for 0 to 13 years and individual ages vary between 39 and 79 (mean: 62.67, standard deviation: 10.96). We used this dataset as an independent test set to validate the results we have obtained on our multiple sound recordings dataset.

III. Methodology

A. Overall Structure of PD Diagnosis System

During this study, data are collected via a physician's examination process. The overall structure of the PD diagnosis system is shown in Fig. 2. When the patient arrives at the hospital, his/her demographic information including gender, age [23], [24], profession, educational status, and a brief health history including the chronic diseases, smoking rate, permanently used drugs, and symptoms of diseases are recorded. Demographic and clinical history information collected in the context of this study are not used in the PD-diagnosis system but only to design a computer-aided data storage system in the

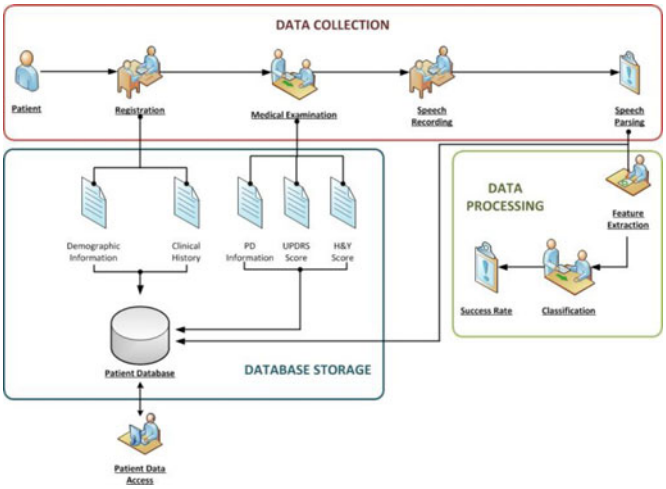

Fig. 2. Overall structure of the system.

hospital. Then the patient is taken to medical examination. During the examination, the physician asks the patient to read or repeat a predefined text including voice samples, write his/her name and draw some loops while conjointly the physician determines an UPDRS score and a stage from Hoehn & Yahr Scale (H&Y). The speech of each patient is recorded during this process. Subsequently, the speech is parsed to be split into voice samples, and time-frequency based features are extracted from the voice samples using Praat acoustic analysis software [25]. The extracted features are fed into a classifier with different cross-validation methods and accuracy, specificity, sensitivity, and MCC evaluation metrics are reported.

B. Feature Extraction

Dysphonia is a well-known speech problem to diagnose and follow the condition of a PWP. Dysphonia leads to reduced loudness, breathiness, roughness, decreased, and exaggerated vocal tremor in voice. These indications can be detected by analyzing various frequencies in voice. In this context, during medical examinations, each subject is asked to read or say predetermined 26 voice samples containing numbers from 1 to 10, four rhymed sentences, nine words in Turkish language along with sustained vowels "a", "o", and "u." To extract features from voice samples, Praat acoustic analysis software is used. A group of 26 linear and time-frequency based features (Table I) are extracted from each voice sample considering the previous works held on this field of study [7], [18].

C. Classification

1) Classification With Leave-One-Subject-Out: The speech datasets collected in previous studies for building noninvasive PD diagnosis typically contain multiple speech recordings per subject, as in the case with our study. Using conventional bootstrapping or leave-one-out validation methods [26], [27] results in biased predictive models by sparing some samples of an individual in the training and some for the testing and creating an artificial overlap between the training and test sets [18]. Instead,

TABLE I
TIME-FREQUENCY-BASED FEATURES EXTRACTED FROM VOICE SAMPLES

Features	Group
Jitter (local) Jitter (local, absolute) Jitter (rap) Jitter (ppq5) Jitter (ddp)	**Frequency parameters**
Number of pulses Number of periods Mean period Standard dev. of period	**Pulse Parameters**
Shimmer (local) Shimmer (local, dB) Shimmer (apq3) Shimmer (apq5) Shimmer (apq11) Shimmer (dda)	**Amplitude parameters**
Fraction of locally unvoiced frames Number of voice breaks Degree of voice breaks	**Voicing Parameters**
Median pitch Mean pitch Standard deviation Minimum pitch Maximum pitch	**Pitch Parameters**
Autocorrelation Noise-to-Harmonic Harmonic -to-Noise	**Harmonicity Parameters**

the proposed classification models in existing studies typically use leave-one-subject-out (LOSO) validation scheme in which all the voice samples of one individual is left out to be used for validation as if it is an unseen individual, and the rest of the samples is used for training. According to the LOSO validation scheme, if the majority of the voice samples of a test individual are classified as PWP, then the individual is classified as positive and otherwise negative.

2) Classification With Summarized Leave-One-Out: We compare the success of conventional LOSO validation scheme with an unbiased LOO method which we called s-LOO (summarized-leave-one-out). According to the s-LOO method, the feature values of 26 voice samples of each subject are summarized using central tendency and dispersion metrics such as mean, median, trimmed mean (10% and 25% removed), standard deviation, interquartile range, mean absolute deviation, and a new form of dataset consisting of N samples is formed where N is the number of subjects. We fed the samples of this dataset to classifiers in groups of two or six metrics using LOO method since multiple voice samples of each individual are degraded to single sample. The groups of six metrics include all the metrics mentioned previously with slight differences in regulations, whereas the groups of two metrics are binary combinations of central tendency and dispersion metrics. By this method, the data are shortened in sample dimension whereas extended in feature dimension.

The aim of summarizing the voice samples of subjects is to decrease the effect of variations between different voice samples of a subject. These variations are originated from the fact that not all the voice samples of a PWP show the dysphonia indications. Thus, labeling all the voice samples of a PWP as positive and

learning the model with this label vector misguide the classifier, which results in a weak prediction model.

3) Evaluation Metrics: The evaluation metrics used to measure the success of classifiers are accuracy, sensitivity, specificity, and MCC. As known, accuracy is the ratio of correctly classified instances to whole instances:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

where *TP* is the number of true positives, *TN* true negatives, *FP* false positives, and *FN* false negatives. Sensitivity and specificity are statistical measures of correctly classified positive and negative instances, respectively:

$$sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$specificity = \frac{TN}{TN + FP}. \tag{3}$$

MCC is a measure that shows the quality of binary classification in machine learning. It is stable even if the class densities are considerably different. MCC is a correlation coefficient between the predicted and observed binary classifications, and gets a value between –1 and +1. The formulation of MCC metric is given as follows:

$$MCC = \frac{TPxTN - FPxFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{4}$$

This coefficient gets the value of +1 when the classifier makes perfect predictions, –1 when the predictions and actual values totally disagree, and 0 when the classification is no better than a random prediction.

4) Statistical Significance of the Results: The statistical significance difference in accuracy between the cross-validation techniques with k-NN and SVM classifiers are tested by McNemar's test [28]. This test is used in dichotomous classification to identify whether two algorithms have the same error rate or not. The number of samples misclassified by both (e_{00}), by the first algorithm but not the second (e_{01}), by the second algorithm but not first (e_{10}) and the number of samples correctly classified by both (e_{11}) are calculated. Then a 2×2 contingency table is created by placing these values. Eventually a chi-square statistic with one degree of freedom is worked out by the formula

$$\frac{(|e_{01} - e_{10} - 1)^2}{e_{01} + e_{10}} \sim X_1^2. \tag{5}$$

If the value of X_1^2 is less than$X_{\alpha,1}^2$, the two algorithms are considered to have the same error rate at significance level α. For $\alpha = 0.05$, $X_{0.05,1}^2 = 3.84$.

IV. Experimental Results

After the normalization process as a preprocessing step so that each feature has a zero mean and standard deviation of one, the features are fed into SVM and k-NN classifiers for PD diagnosis. For the k-NN classifier, Euclidean distance metric and k parameter of 1, 3, 5, and 7 are used. For SVM application, LIBSVM [29] package is used with a linear and radial basis function (RBF) kernels along with cost value (c) parameter of 10 and kernel width (g) of 0.005.

TABLE II
k-NN Classification Accuracies Using LOSO and s-LOO

k	Cross Validation	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
1	**LOSO**	53.37	.0007	49.62	57.12
	s-LOO (1-4)	42.50	.0015	30.00	55.00
	s-LOO (2-5)	52.50	.0005	45.00	60.00
	s-LOO (3-6)	50.00	.0000	55.00	45.00
	s-LOO (all)	55.00	.1000	55.00	55.00
3	**LOSO**	54.04	.0008	53.27	54.81
	s-LOO (1-4)	55.00	.1021	45.00	65.00
	s-LOO (2-5)	60.00	.2010	55.00	65.00
	s-LOO (3-6)	42.50	-.0015	55.00	30.00
	s-LOO (all)	55.00	.1000	55.00	55.00
5	**LOSO**	54.42	.0009	53.65	55.19
	s-LOO (1-4)	55.00	.1021	45.00	65.00
	s-LOO (2-5)	57.50	.1517	65.00	50.00
	s-LOO (3-6)	50.00	.0000	70.00	30.00
	s-LOO (all)	55.00	.1048	70.00	40.00
7	**LOSO**	53.94	.0008	54.04	53.85
	s-LOO (1-4)	**65.00**	**.3062**	55.00	75.00
	s-LOO (2-5)	62.50	.2503	60.00	65.00
	s-LOO (3-6)	42.50	-.0017	65.00	20.00
	s-LOO (all)	57.50	.1517	65.00	50.00

Central tendency metrics: 1: mean **2:** median **3:** trimmed mean (25% removed).
Dispersion metrics: 4: standard deviation **5:** mean absolute deviation **6:** interquartile range.

The results obtained using various subsets formed from the original dataset using different k-NN and SVM options are compared. The k-NN classification accuracies using LOSO and s-LOO with different number of nearest neighbors (k parameter values) are shown in Table II. As seen from the results, almost a random prediction is obtained using the samples with conventional LOSO cross-validation method for all values of k parameter (see the MCC values). The highest MCC and overall accuracy obtained are 0.3062 and 65.00%, respectively, with s-LOO method by summarizing the data using mean as the central tendency and standard deviation as the dispersion metric ($k = 1$). Sensitivity is also another important evaluation metric in the field of biomedical science since the early diagnosis of a disease increases the chance of treatment and helps to prevent the symptoms from becoming worse. Summarizing the data using trimmed mean (25% removed)-interquartile range pair and all the central tendency-dispersion metrics with k parameter of 5 produced the highest sensitivity (70.00%).

The SVM classifier produced higher accuracies than the k-NN classifier, as can be seen in Table III. The highest accuracy (77.50%) is obtained with s-LOO method by summarizing the data using the mean-standard deviation binary combination of central tendency and dispersion metrics as in the case of k-NN classification results. This model also produced the highest MCC, sensitivity, and specificity values. The SVM classifier with linear kernel using LOSO method also produced almost a random prediction (MCC = 0.0006) whereas RBF kernel

TABLE III
SVM CLASSIFICATION ACCURACIES USING LOSO AND S-LOO

kernel	Cross Validation	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
Linear	LOSO	52.50	.0006	52.50	52.50
	s-LOO (1-4)	**77.50**	**.5507**	80.00	75.00
	s-LOO (2-5)	70.00	.4082	80.00	60.00
	s-LOO (3-6)	60.00	.2000	65.00	45.00
	s-LOO (all)	67.50	.3504	70.00	65.00
RBF	LOSO	55.00	.1005	60.00	50.00
	s-LOO (1-4)	65.00	.3015	60.00	70.00
	s-LOO (2-5)	70.00	.4000	70.00	70.00
	s-LOO (3-6)	72.50	.4506	70.00	75.00
	s-LOO (all)	65.00	.3015	70.00	60.00

Central tendency metrics: 1: mean **2:** median **3:** trimmed mean (25% removed).
Dispersion metrics: 4: standard deviation **5:** mean absolute deviation **6:** interquartile range.

TABLE IV
MCNEMAR'S TEST BETWEEN SVM PREDICTIONS OF LOSO AND S-LOO WITH MEAN-STANDARD DEVIATION

		LOSO	
		Misclassified	Correct
s-LOO	Misclassified	8	1
	Correct	11	20
$X_1^2 = 6.75$ α=0.05			

TABLE V
AVERAGE AND BEST RESULTS OF 1000 RUNS OF SELECTING A RANDOM VOICE SAMPLES FROM EACH INDIVIDUAL

Classifier	Parameter	Result	Acc. (%)	MCC	Sens. (%)	Spec. (%)
k-NN	1	average	50.61	.0124	52.71	48.52
		best	**82.50**	**.6580**	85.00	80.00
	3	average	49.49	-.0102	54.61	44.37
		best	77.50	.5563	70.00	85.00
	5	average	48.52	-.0298	56.12	40.93
		best	75.00	.5103	65.00	85.00
	7	average	48.12	-.0383	57.18	39.07
		best	77.50	.5563	85.00	70.00
SVM	linear kernel	average	52.06	.0416	54.92	49.22
		best	**85.00**	**.7035**	80.00	90.00
	RBF kernel	average	46.91	-.0618	49.21	44.62
		best	80.00	.6030	85.00	75.00

produced a better prediction model (MCC = 0.1005). It is seen that all the models with s-LOO are more successful than LOSO in distinguishing the patients with PD from healthy subjects. The results also indicate that SVM gave more stable results than the k-NN classifier. The significance of the difference in accuracy between the linear SVMs classifier using s-LOO with mean-standard deviation and conventional LOSO methods is assessed by McNemar's test (Table IV). McNemar's test revealed that the accuracy of s-LOO with mean-standard deviation is higher than of LOSO at significance level 0.05 ($X_1^2 = 6.75$).

In order to examine the success of our s-LOO method from a different point of view, we created new subsets of data by selecting a random voice sample of each individual. These randomly created subsets are then fed to classifiers. We repeated this process 1000 times and reported the average and best results. As seen in Table V, the average accuracy obtained using random selection can only get as high as 52.06% with linear kernel SVMs. The average MCC is only 0.0416 meaning that shuffling the recordings indeed results in complete failure of the classifier and the original s-LOO accuracy was not just accidental. Among the 1000 runs of creating random pseudo-subjects out of the whole set of speech recordings, the statistical significance of the original accuracy of s-LOO has been found to be very high: In terms of accuracy, $p = 0.004$ and in terms of MCC, $p = 0.006$.

TABLE VI
SVM CLASSIFICATION ACCURACIES OF INDIVIDUAL VOICE SAMPLES THAT YIELDED THE BEST RESULTS

Voice Sample	Kernel Type	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
Sustained "o"	Linear Kernel	**72.50**	**.4506**	70.00	75.00
	RBF Kernel	50.00	.0000	55.00	45.00
"four"	Linear Kernel	72.50	.4506	75.00	70.00
	RBF Kernel	**75.00**	**.5000**	75.00	75.00

Another experiment we performed is feeding each voice sample of the subjects individually to the classifier one at a time. We reported the voice samples that yielded best results in Table VI. The LOO cross-validation method is used since there is only one record of each subject from each voice sample in the created subdatasets. The best results are achieved using the features of sustained "o" and "four" voice samples which show that in case of a temporal constraint, these two words can be used to detect the voice indicators of PD. The chi-square statistics indicated that the error rates of the proposed approach LOSO and the individual voice samples that yielded the best results have the same error rates at significance level 0.05. This comparison reveals that even the best individual voice samples, which have bias since initially classification algorithms are run for each voice sample and the ones that generated best classification accuracy are considered, do not yield significantly better prediction models than s-LOO. It must also be noted that the voice samples other than sustained "o" and "four" produce poor classification accuracies and MCC values when compared to s-LOO method.

The experimental results point out that collecting as many voice samples as possible from patients and summarizing the extracted features of each voice sample with central tendency and dispersion metrics increases the success of the diagnostic system. In Fig. 3, the change in classification accuracy and MCC of s-LOO with mean-standard deviation and LOSO methods according to the increasing number of voice samples is shown. As seen from the figure, using the mean and standard deviation of voice samples ends up with higher classification accuracy and MCC values. The best result yielded using the first ten voice samples; however, it is not a reliable result since some other ten voice sample combinations may derive higher accuracy and MCC values. However, in general it can be said that using more samples increases both accuracy and MCC.

The LOSO and s-LOO methods are applied on the independent test set for the validation of the obtained results. For this purpose, the k-NN and SVM classifiers are trained using the sustained vowel voice samples of multiple sound recordings

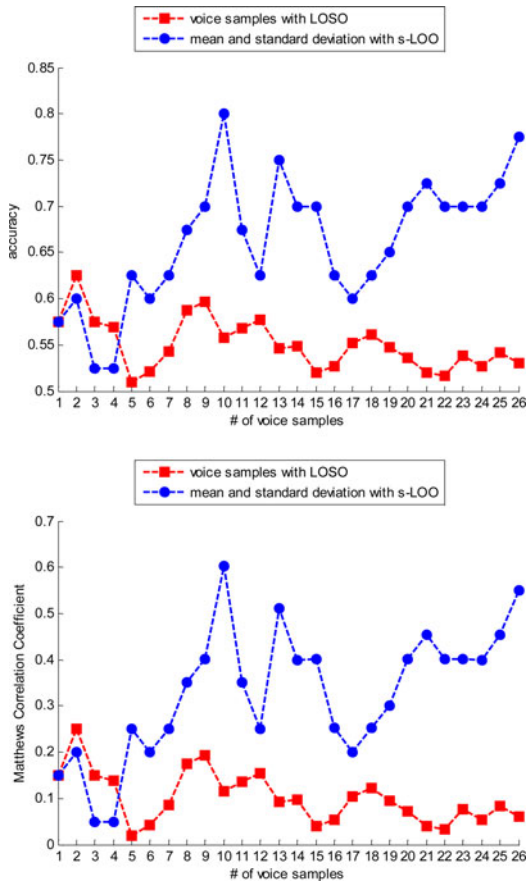

Fig. 3. SVM model using LOSO versus s-LOO with mean-standard deviation (left) classification accuracy (right) MCC.

TABLE VII
k-NN CLASSIFICATION ACCURACIES (%) ON INDEPENDENT TEST SET

k	LOSO	s-LOO (1-4)	s-LOO (2-5)	s-LOO (3-6)	s-LOO (all)
1	50.60	60.71	67.86	57.14	64.29
3	48.21	57.14	53.57	71.43	67.86
5	53.57	**78.57**	67.86	**78.57**	71.43
7	59.52	75.00	71.43	75.00	67.86

Central tendency metrics: 1: mean **2:** median **3:** trimmed mean (25% removed)
Dispersion metrics: 4: standard deviation **5:** mean absolute deviation **6:** interquartile range

TABLE VIII
SVM CLASSIFICATION ACCURACIES (%) ON INDEPENDENT TEST SET

kernel	LOSO	s-LOO (1-4)	s-LOO (2-5)	s-LOO (3-6)	s-LOO (all)
linear	58.33	60.71	67.86	67.86	60.71
RBF	68.45	**82.14**	71.43	71.43	75.00

Central tendency metrics: 1: mean **2:** median **3:** trimmed mean (25% removed)

dataset and the obtained models are tested on the independent dataset. The k-NN and SVM classification accuracies are shown in Tables VII and VIII, respectively. It is seen that, in parallel to the results we have observed on the multiple sound recordings dataset, s-LOO produced predictive models with higher classification accuracies than conventional LOSO. Likewise, the highest overall accuracies of both k-NN and SVM classifiers are obtained with s-LOO method by summarizing the data using mean as the central tendency and standard deviation as the dispersion metric.

V. CONCLUSION

Due to the recent interest in speech pattern analysis applications of PD for building predictive telediagnosis and telemonitoring models, we have collected a wide variety of voice samples and various sound types, including sustained vowels, words, and sentences compiled from a set of speaking exercises for PWP. Apart from the findings presented in our study, it presents the opportunity to explore the international applicability and validity of the previously built models and is expected to attract the biomedical signal processing and machine learning societies by presenting a separate test dataset for the models solely based on the PD dataset previously published by [7].

As a result of the analysis of our dataset, in parallel to the results reported in the literature, sustained vowels have been found to carry more PD-discriminative information than the isolated words and short sentences do. To evaluate which and how well the central tendency and dispersion metrics (among mean, median, trimmed mean, standard deviation, interquartile range, and mean absolute deviation) of a subject serve as good representatives of all his recordings, we have tried several combinations of these metrics and have found that representing the samples of a subject with the classical mean and standard deviation improves generalization of the predictive model. This type of representation is shown to be more effective than using each voice recording of each subject as an independent data sample. Using the mean and standard deviation of the vocal features as a summarizing representation for multiple recordings of each subject has been shown to be an efficient strategy in building such predictive models.

ACKNOWLEDGMENT

The study has been approved by the Istanbul University Research Ethics Committee. The authors would like to thank Dr. A. Tsanas and Dr. M. Little from the University of Oxford for helpful discussions.

REFERENCES

[1] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol. Neurosurgery Psychiatry*, vol. 79, no. 4, pp. 368–376, 2007.
[2] J. W. Langston, "Parkinson's disease: Current and future challenges," *NeuroToxicology*, vol. 23, no. 4-5, pp. 443–450, 2002.
[3] S. B. O'Sullivan and T. J. Schmitz, "Parkinson disease," in *Physical Rehabilitation*, 5th ed. Philadelphia, PA, USA: F. A. Davis Company, 2007, pp. 856–894.
[4] Parkinson Derneği. (2011). [Online]. Available: http://www.parkinsondernegi.org/Icerik.aspx?Page=parkinsonnedir&ID=5
[5] L. M. de Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *Lancet Neurol.*, vol. 5, no. 6, pp. 525–535, 2006.
[6] N. Singh, V. Pillay, and Y. E. Choonara, "Advances in the treatment of Parkinson's disease," *Prog. Neurobiol.*, vol. 81, no. 1, pp. 29–44, 2007.
[7] 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. 1010–1022, Apr. 2009.
[8] National Collaborating Centre for Chronic Conditions, *Parkinson's Disease*, London, U.K.: Royal College of Physicians, 2006.

[9] L. Cunningham, S. Mason, C. Nugent, G. Moore, D. Finlay, and D. Craig, "Home-based monitoring and assessment of Parkinson's disease," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 1, pp. 47–53, Jan. 2011.

[10] G. Rigas, A. Tzallas, M. Tsipouras, P. Bougia, E. Tripoliti, D. F. D. Baga, S. Tsouli, and S. Konitsiotis, "Assessment of tremor activity in the Parkinson's disease using a set of wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 478–487, May 2012.

[11] S. Marino, R. Ciurleo, G. Lorenzo, M. Barresi, S. De Salvo, S. Giacoppo, A. Bramanti, P. Lanzafame, and P. Bramanti, "Magnetic resonance imaging markers for early diagnosis of Parkinson's disease," *Neural Regeneration Res.*, vol. 7, no. 8, pp. 611–619, 2012.

[12] Z. Dastgheib, B. Lithgow, and Z. Moussavi, "Diagnosis of Parkinson's disease using electrovestibulography," *Med. Biol. Eng. Comput.*, vol. 50, no. 3, pp. 483–491, 2012.

[13] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomed. Eng. Online*, vol. 6, no. 23, 2007, doi: 10.1186/1475-925X-6-23.

[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] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramige, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *J. Royal Society Interface*, vol. 8, pp. 842–855, 2011.

[16] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, Apr. 2010.

[17] A. Tsanas, M. A. Little, P. E. McSharry, B. K. Scanlon, and S. Papapetropoulos, "Statistical analysis and mapping of the Unified Parkinson's Disease Rating Scale to Hoehn and Yahr staging," *Parkinsonism Relat. Disord.*, vol. 18, no. 5, pp. 697–699, 2012.

[18] C. O. Sakar and O. Kursun, "Telediagnosis of Parkinson's disease using measurements of dysphonia," *J. Med. Syst.*, vol. 34, no. 4, pp. 591–599, 2010.

[19] O. Kursun, E. Gumus, A. Sertbas, and O. Favorov, "Selection of vocal features for Parkinson's disease diagnosis," *Int. J. Data Mining Bioinf.*, vol. 6, no. 2, pp. 144–161, 2012.

[20] I. Bhattacharya and M. Bhatia, "SVM classification to distinguish Parkinson disease patients," in *A2CWiC '10 Proc. 1st Amrita ACM-W Celebration on Women in Computing in India*, New Delhi, 2010, doi: 10.1145/1858378.1858392.

[21] L. O. Ramig, S. Sapir, C. Fox, and S. Countryman, "Changes in vocal loudness following intensive voice treatment (LSVT®) in individuals with Parkinson's disease: A comparison with untreated patients and normal age-matched controls," *Mov. Disord.*, vol. 16, no. 1, pp. 79–83, 2001.

[22] H. Apaydın, S. Özekmekçi, S. Oğuz, and İ. Zileli, *Parkinson Hastalığı, Hasta ve Yakınları için El Kitabı*. İstanbul, Turkey: İstanbul Üniversitesi Yayınları, 2008.

[23] J. A. Obeso, M. C. Rodriguez-Oroz, C. G. Goetz, C. Marin, J. H. Kordower, M. Rodriguez, E. C. Hirsch, M. Farrer, A. H. V. Schapira, and G. Halliday, "Missing pieces in the Parkinson's disease puzzle," *Nature Med.*, vol. 16, no. 6, pp. 653–661, 2012.

[24] B. Post, M. P. Merkus, and R. J. de Haan, "Prognostic factors for the progression of Parkinson's disease: A systematic review," *Mov. Disord.*, vol. 22, no. 13, pp. 1839–1851, 2007.

[25] P. Boersma and D. Weenink. (2012). Praat: Doing phonetics by computer [Online]. Available: http://www.praat.org/. [Accessed 29 March 2012]

[26] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Stat.*, vol. 7, no. 1, pp. 1–26, 1979.

[27] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1371–1382, 2003.

[28] E. Alpaydin, *Introduction to Machine Learning*. London, U.K.: MIT Press, 2010.

[29] C. W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

Authors' photographs and biographies not available at the time of publication.