Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease

Athanasios Tsanas*, Max A. Little, Patrick E. McSharry, *Senior Member, IEEE*, Jennifer Spielman, and Lorraine O. Ramig

***Abstract*—There has been considerable recent research into the connection between Parkinson's disease (PD) and speech impairment. Recently, a wide range of speech signal processing algorithms (dysphonia measures) aiming to predict PD symptom severity using speech signals have been introduced. In this paper, we test how accurately these novel algorithms can be used to discriminate PD subjects from healthy controls. In total, we compute 132 dysphonia measures from sustained vowels. Then, we select four parsimonious subsets of these dysphonia measures using four feature selection algorithms, and map these feature subsets to a binary classification response using two statistical classifiers: random forests and support vector machines. We use an existing database consisting of 263 samples from 43 subjects, and demonstrate that these new dysphonia measures can outperform state-of-the-art results, reaching almost 99% overall classification accuracy using only ten dysphonia features. We find that some of the recently proposed dysphonia measures complement existing algorithms in maximizing the ability of the classifiers to discriminate healthy controls from PD subjects. We see these results as an important step toward noninvasive diagnostic decision support in PD.**

***Index Terms*—Decision support tool, feature selection (FS), Parkinson's disease (PD), nonlinear speech signal processing, random forests (RF), support vector machines (SVM).**

Manuscript received July 18, 2011; revised October 31, 2011; accepted. December 26, 2011. Date of publication January 9, 2012; date of current version April 20, 2012. This work was supported in part by the National Institutes of Health (NIH) under Grant R01 DC1150 (National Institutes of Deafness and Other Communication Disorders). The work of A. Tsanas was supported by the Engineering and Physical Sciences Research Council (EPSRC), U.K., and by Intel Corporation, Santa Clara, CA. The work of M. A. Little was supported by the Welcome Trust under Grant WT090651MF. *Asterisk indicates corresponding author.*

*A. Tsanas is with the Oxford Centre for Industrial and Applied Mathematics (OCIAM), Mathematical Institute, University of Oxford, Oxford OX1 3LB, U.K., and also with the Systems Analysis Modelling and Prediction (SAMP) Group, Department of Engineering Science, University of Oxford, Oxford OX1 3PG, U.K. (e-mail: tsanas@maths.ox.ac.uk).

M. A. Little is with the Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139 USA, and also with the Department of Physics, University of Oxford, Oxford OX1 3RH, U.K. (e-mail: maxl@mit.edu).

P. E. McSharry is with the Smith School of Enterprise and the Environment, University of Oxford, Oxford OX1 2BQ, U.K., and also with the Oxford Centre for Industrial and Applied Mathematics, University of Oxford, Oxford OX1 3LB, U.K. (e-mail: patrick@mcsharry.net).

J. Spielman and L. O. Ramig are with the Department of Speech, Language, and Hearing Science, University of Colorado, Boulder, CO 80309-0001 USA, and also with the National Center for Voice and Speech, Denver , CO 80204 USA (e-mail: Jennifer.Spielman@Colorado.EDU; Lorraine.Ramig@colorado.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

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/TBME.2012.2183367

I. Introduction

NEUROLOGICAL disorders affect people's lives at an epidemic rate worldwide. Parkinson's disease (PD) is one of the most common neurodegenerative disorders with an incidence rate of approximately 20/100 000 [1] and a prevalence rate exceeding 100/100 000 [2]. Moreover, these statistics might underestimate the problem because PD diagnosis is complicated [3]. Given that age is the single most important factor for PD and the fact that the population is growing older, these figures could further increase in the not too distant future [4].

Identifying the causes of PD onset remains elusive, although genetic and environmental factors may be implicated [1]; hence, the disease is often referred to as *idiopathic*. In those cases where particular factors can be identified that cause PD-like symptoms (for example drugs), the disease is termed *Parkinsonism*. PD symptoms include tremor, rigidity and loss of muscle control in general, as well as cognitive impairment.

The difficulty in *reliable* PD diagnosis has inspired researchers to develop *decision support tools* relying on algorithms aiming to differentiate healthy controls from people with Parkinson's (PWP) [5]–[7]. Although this binary discrimination approach does not form a *differential diagnosis* (a differential diagnostic tool should be able to distinguish PD subjects amongst a variety of disorders that present PD-like symptoms), it is a promising first step toward that long-term goal.

Research has shown that *speech* may be a useful signal for discriminating PWP from healthy controls [5], [7], on the basis of clinical evidence which suggests that the vast majority of PWP typically exhibit some form of vocal disorder [8]. In fact, vocal impairment may be amongst the earliest prodromal PD symptoms, detectable up to five years prior to clinical diagnosis [9]. In our own research, we have also presented strong evidence linking speech to average Parkinson's disease symptom severity [5], [10]–[13]. Collectively, these findings reinforce the notion that speech may reflect disease status, after appropriate processing of the recorded speech signals.

The range of symptoms present in speech includes reduced loudness, increased vocal tremor, and breathiness (noise). Vocal impairment relevant to PD is described as *dysphonia* (inability to produce normal vocal sounds) and *dysarthria* (difficulty in pronouncing words). We refer to Baken and Orlikoff [14] for a more detailed description of speech disorders. The extent of vocal impairment is typically assessed using *sustained vowel* phonations, or *running speech*. Although it can be argued that some of the vocal deficiencies in running speech (such as combinations of consonants and vowels) might not be captured

by the use of sustained vowels, the analysis of running speech is more complex due to articulatory and other linguistic confounds [15], [16]. Therefore, the use of sustained vowels, where the speaker is requested to sustain phonation for as long as possible, attempting to maintain steady frequency and amplitude at a comfortable level, is commonplace in clinical practice [15]. Research has shown that the sustained vowel "ahh. . ." is sufficient for many voice assessment applications [15], including PD status prediction [5] and average PD symptom monitoring [10], [11].

The study of speech disorders in general and in the context of PD, in particular, has prompted the development of many speech signal processing algorithms (henceforth *dysphonia measures*), for example, see [5], [7], [11], [15], and references therein. In [5], it was shown that the most commonly used speech signal processing algorithms could discriminate PWP from healthy controls with approximately 90% overall classification accuracy, using four dysphonia features. That study included traditional measurement algorithms focusing on fundamental frequency perturbation (*jitter measures*), amplitude perturbation (*shimmer measures*), and signal-to-noise ratios (SNRs) (harmonics-to-noise ratio measures). Moreover, that study included three novel nonlinear dysphonia measures, complementing the classical measures (see Section II-A).

Subsequently, the dysphonia measures of [5] were applied to the study of the related problem of mapping speech impairment to average PD symptom severity [10]. Very recently, additional nonlinear dysphonia measures have been proposed for that application [11], which (coupled with some classical algorithms) significantly improved on previous results [10]. Hence, we hypothesized that applying the dysphonia measures of [11] to the problem of discriminating PWP from healthy controls might bring additional insight, and improved results [5].

II. Data

The National Center for Voice and Speech (NCVS) database comprises 263 phonations from 43 subjects (17 females and 26 males, 10 healthy controls, and 33 PWP), an extension of the database used in [5] (the extended database includes all the voice recordings from the earlier study). The ten healthy controls (four males and six females), had an age range of 46–72 years with (mean $\pm$ standard deviation) 61 ± 8.6 years, and we processed 61 healthy phonations. The 33 PWP (22 males and 11 females), had an age range of 48–85 (67.2 ± 9.3), time since diagnosis 0 to 28 years (5.8 ± 6.3); there are 202 PD phonations. This database comprises six or seven sustained vowel "ahh. . ." phonations from each speaker, recorded at a comfortable frequency and amplitude.

The phonations were recorded in an IAC sound-treated booth with a head-mounted microphone (AKG C420), which was placed at 8-cm distance from the subject's mouth. The voice signals were sampled at 44.1 kHz with 16 bits resolution, and were recorded directly to computer using CSL 4300B hardware (Kay Elemetrics).

TABLE I
Breakdown of the 132 Dysphonia Measures Used in This Study

Family of dysphonia measures	Brief description	Number of measures
Jitter variants	F0 perturbation	30
Shimmer variants	Amplitude perturbation	21
Harmonics to noise ratio (HNR) and noise to harmonics ratio (NHR)	Signal to noise, and noise to signal ratios	4
Glottis quotient (GQ)	Vocal fold cycle duration changes	3
Recurrence period density entropy (RPDE)	Uncertainty in estimation of fundamental frequency	1
Detrended fluctuation analysis (DFA)	Stochastic self-similarity of turbulent noise	1
Pitch period entropy (PPE)	Inefficiency of F0 control	1
Glottal to noise excitation (GNE)	Extent of noise in speech using energy and nonlinear energy concepts	6
Vocal fold excitation ratio (VFER)	Extent of noise in speech using energy, nonlinear energy, and entropy concepts	9
Empirical mode decomposition excitation ratio (EMD-ER)	Signal to noise ratios using EMD-based energy, nonlinear energy and entropy	6
Mel Frequency Cepstral Coefficients (MFCC)	Amplitude and spectral fluctuations	42
F0-related measures	Summary statistics of F0, Differences from expected F0 in age- and sex- matched controls, variations in F0	8

Algorithmic expressions for the 132 measures summarized here are described in detail in Tsanas *et al*. [11]. F0 refers to fundamental frequency estimates.

III. Methods

The aim of this study is to analyze the speech signals, extract *features*, and to attempt to map these features to the *response* (PD versus healthy control).

A. Extracting Features From the Speech Signals

We use the dysphonia measures rigorously defined in [11]. The rationale, background, and algorithms used to compute these features are also explained in detail in that paper. Here, we summarize these algorithms. For convenience, Table I lists the extracted features, grouped together into algorithmic "families" of features that share common attributes, along with a brief description of the properties of the speech signals that these algorithms aim to characterize.

Typical examples of features are jitter and shimmer [14], [15]. The motivation for these features is that the vocal fold vibration pattern is nearly periodic in healthy voices, whereas this periodic pattern is considerably disturbed in pathological cases [15]. Therefore, PWP are expected to exhibit relatively large values of jitter and shimmer compared to healthy controls. Different studies use slightly different definitions of jitter and shimmer, for example, by normalizing the measure over a different range of *vocal fold cycles* (time interval between successive vocal fold collisions). For that reason, here we investigate many variations of these algorithms which we collectively refer to as *jitter* and *shimmer variants* [11].

Building on the concept of irregular vibration of the vocal folds, earlier studies have proposed the recurrence period density entropy (RPDE), the pitch-period entropy (PPE), the glottis quotient (GQ), and F0-related measures [5], [11]. GQ attempts to detect vocal fold *cycle durations* [17]. Then, we work directly on the variations of the estimated cycle durations to obtain the GQ measures. RPDE quantifies the *uncertainty* in estimation of the vocal fold cycle duration using the information theoretic concept of entropy. PPE uses the log-transformed linear prediction residual of the fundamental frequency in order to smooth normal *vibrato* (normal, small, periodic perturbations of the vocal fold cycle durations which are present in both healthy and PD voices [15]), and measures the impaired control of fundamental frequency (F0) during sustained phonation. The F0-related measures (such as the standard deviation of the F0 estimates) include the difference in the measured F0 with the expected, healthy F0 in the population for age- and gender-matched controls [15].

The second general family of dysphonia measures quantifies noise, or produces a SNR estimate. The physiological motivation for these measures is that pathological voices exhibit increased aeroacoustic noise because of the creation of excessive turbulence due to incomplete vocal fold closure. Such measures include the harmonic-to-noise ratio (HNR), detrended fluctuation analysis (DFA), glottal to noise excitation (GNE), vocal fold excitation ratio (VFER), and empirical mode decomposition excitation ratio (EMD-ER). GNE and VFER analyze the full frequency range of the signal in bands of 500 Hz [11]. Additionally, we have created SNR measures using energy, nonlinear energy (Teager–Kaiser energy operator) and entropy concepts whereby the frequencies below 2.5 kHz are treated as "signal", and everything above 2.5 kHz is treated as "noise" [11]. EMD-ER has a similar justification: the Hilbert–Huang transform [18] decomposes the original signal into components, where the initial components are the high-frequency constituents (in practice equivalent to noise), and the later components constitute useful information (actual signal).

Finally, *mel-frequency cepstral coefficients* (MFCC) have long been used in speaker identification and recognition applications, but have shown promise in recent biomedical voice assessments [11], [19], [20]. They are aimed at detecting subtle changes in the motion of the articulators (tongue, lips), which are known to be affected in PD [8].

Overall, applying the 132 dysphonia measures to the 263 NCVS speech signals, gave rise to a 263×132 *feature matrix*. There were no missing entries in the feature matrix.

B. Preliminary Statistical Survey of Dysphonia Features

In order to gain a preliminary understanding of the statistical properties of the features, we computed the Pearson correlation coefficient and the mutual information $I(\boldsymbol{x},\boldsymbol{y})$, where the vector $\boldsymbol{x}$ contains the values of a single feature for all phonations, and $\boldsymbol{y}$ is the associated response. As in [11], we normalize $I(\boldsymbol{x},\boldsymbol{y})$ by dividing through $I(\boldsymbol{y},\boldsymbol{y})$ for presentation purposes. The larger the value of the normalized mutual information, the stronger the statistical association between the feature and the response. We used the KDE Toolbox by Ihler and Mandel for the computation of the mutual information [21]. The mutual information is computed via the evaluation of the marginal entropies $H(\boldsymbol{x})$, $H(\boldsymbol{y})$ and the joint entropy $H(\boldsymbol{x},\boldsymbol{y})$. The entropies are computed by evaluating the mean log-likelihood of the density estimates (the densities are computed using kernel density estimation with Gaussian kernels) [21].

C. Feature Selection

With the large number of dysphonia features of this study, we cannot expect the feature space to be uniformly populated by only 263 phonations, and the risk of overfitting arises. Many classification algorithms are fairly robust to the inclusion of potentially noisy or irrelevant features, and their predictive power may or may not be *severely* affected; however, reducing the number of features often improves the model's predictive power for hold-out data. A reduced feature subset also facilitates inference, enabling one to gain insights into the problem via analysis of the most predictive features [22], [23].

Exhaustive search through all possible feature subsets is computationally intractable, a problem which has led to the development of *feature selection* (FS) *algorithms* which offer a rapid, principled approach to reduction of the number of features. FS is a topic of extensive research, and we refer to Guyon *et al*. [23] for further details.

Here, we have compared four efficient FS algorithms: 1) least absolute shrinkage and selection operator (LASSO) [24], 2) minimum redundancy maximum relevance (mRMR) [25], 3) RELIEF [26], and 4) local learning-based feature selection (LLBFS) [27]. LASSO penalizes the absolute value of the coefficients in a linear regression setting; this leads to some coefficients that are shrunk to zero, which effectively means that the features associated with those coefficients are eliminated. The LASSO has been shown to have *oracle properties* (correctly identifying *all* the "true" features contributing toward predicting the response) in *sparse* settings when the features are not highly correlated [28]. However, when the features are correlated, some noisy features (not contributing toward predicting the response) may still be selected [29]. Moreover, some useful features toward predicting the response amongst the correlated features may be discarded [22]. The mRMR algorithm uses a heuristic criterion to set a tradeoff between maximizing *relevance* (association strength of features with the response) and minimizing *redundancy* (association strength between pairs of features). It is a *greedy* algorithm (selecting one feature at a time), which takes into account only pairwise redundancies and neglects *complementarity* (*joint* association of features toward predicting the response). RELIEF is a *feature-weighting algorithm*, which promotes features that contribute to the separation of samples from different classes. It is conceptually related to margin maximization algorithms, and has been linked to the k-nearest-neighbor classifier [30]. Contrary to mRMR, RELIEF uses complementarity as an inherent part of the FS process. Finally, LLBFS aims to decompose the intractable, exhaustive combinatorial problem of FS into a set of locally linear problems through local learning. The original features are assigned feature weights that denote their importance to the classification

problem, and the features with the maximal weights are then selected. LLBFS was conceived as an extension of RELIEF and relies on kernel density estimation and margin maximization concepts [27]. Overall, all four FS algorithms have shown promising results in machine learning applications over a wide range of different applications.

The feature subsets were selected using a cross-validation (CV) approach (see Section III-E), using only the training data at each CV iteration. We repeated the CV process a total of ten times, where each time the M features ($M = 132$) for each FS algorithm appear in descending order of selection. Ideally, this feature ordering would be identical for all ten CV iterations, but in practice it is not. Hence, we need to have a strategy to select the features that appeared most often under each of the FS algorithms, to identify four feature subsets, one subset for each FS algorithm. Specifically, for each FS algorithm, we create an empty set S which will contain the indices of the features selected, and apply the following voting scheme. Feature indices are incrementally included, one at a time, in S. For each step K (K is a scalar taking values $1, \ldots, M$), we find the indices corresponding to the features selected in the $1, \ldots, K$ search steps for all the ten CV repetitions. Then, we select the index which appears most frequently amongst these $10 \times K$ elements and which is also not already included in S. This index is now included as the Kth element in S. Ties are resolved by including the lowest index number. This entire process is repeated for each of the four FS algorithms. There is one final implementation issue that we need to address: contrary to the other three FS algorithms, LASSO may remove features in subsequent stages during its incremental FS search. Therefore, for LASSO, we repeated the tenfold CV process independently for each Kth step, interrogating the algorithm to provide the best-K features prior to the voting scheme explained before.

Once the final selected feature subset S was decided for each FS algorithm, these features were input into the classifier in the subsequent mapping phase to obtain the final healthy/PD predictions from the dysphonia measures.

D. Mapping Selected Dysphonia Features to the Response

The preliminary correlation analysis of the features against the response presented before provides an indication of the association strength of each feature with the response. However, ultimately our aim is to develop a functional relationship $f(\boldsymbol{X}) = \boldsymbol{y}$, which maps the dysphonia features $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$, where M is the number of features, to the response $\boldsymbol{y}$. That is, we need a *binary classifier* that will use the dysphonia measures to discriminate healthy controls from PWP.

We compared two widely used statistical machine-learning algorithms here: random forests (RF), and support vector machines (SVM) [22]. RF is an *ensemble technique*, weighting the output of a large number of *tree-structured* prediction functions f (we used 500 trees). RF has a single tuning parameter: the number of features over which to search to construct each branch of each tree. However, this classifier has been found to be very robust to the choice of this parameter [32]. Following the suggestion of Breiman [32], we used the default setting (the square root of the number of input features), but also compared the results using half this default number (i.e. the square root of the number of input features, divided by two), and double this number (i.e. the square root of the number of input features, multiplied by two).

SVMs attempt to construct an optimal *separating hyperplane* in the feature space, between the two classes in this binary decision problem by maximizing a *geometric margin* between points from the two classes. In practical applications, data often cannot be *linearly* separated; in those cases, SVMs can use the *kernel trick* to transform the data into a higher dimensional space, and construct the separating hyperplane in that space [22]. There is extensive research, beyond the scope of this study, on how to work with nonlinearly separable data (see [22] and references therein). In general, this classifier requires the specification of some internal parameters, and SVMs are known to be particularly sensitive to the values of these parameters [22]. Here, we used the LIBSVM implementation [33] and followed the suggestions of the developers of that implementation [34]: we linearly scaled each of the input features to lie in the range [–1, 1], and used a Gaussian, radial basis function kernel. The determination of the optimal values of the kernel parameter γ and the penalty parameter C was decided using a *grid search* of possible values. We selected the pair (C,γ) that gave the lowest CV misclassification error (see Section III-E for details of CV scheme). Specifically, we searched over the grid (C,γ) defined by the product of the sets $C = [2^{-5}, 2^{-13}, \ldots, 2^{15}]$, and $\gamma = [2^{-15}, 2^{-13}, \ldots, 2^{3}]$. Once the optimal parameter pair (C,γ) was determined, we trained and tested the classifier using these parameters.

E. Classifier Validation

Validation in this context aims at an estimate of the *generalization* performance of the classification based on the dysphonia features, when presented with novel, previously unseen data. The tacit statistical assumption is that the new, unseen data will have a similar joint distribution to the data used to train the classifier. Most studies achieve this validation using either CV or bootstrap techniques [22].

In this study, we used a tenfold CV scheme, where the original data (263 phonations) were split into two subsets: a training subset consisting of 90% of the data (237 phonations), and a testing subset consisting of 10% of the data (26 phonations). The process was repeated a total of 100 times, where in each repetition the original dataset was randomly permuted prior to splitting into training and testing subsets. On each repetition, we computed the mean absolute classification error $\mathrm{MAE} = 1/N \sum_{i \in Q} |\hat{y}_i - y_i|$, where $\hat{y}_i$ is the predicted response, y_i is the actual response for each ith entry in the training or testing subset, N is the number of phonations in the training or testing subset, and Q contains the indices of that set. Errors over the 100 CV repetitions were averaged. Then, the performance of the model is $(1 - \mathrm{MAE}) \cdot 100\%$.

TABLE II
STATISTICAL ANALYSIS OF THE DYSPHONIA FEATURES

Dysphonia measure	Correlation coefficient	Normalized mutual information
$VFER_{entropy}$	**-0.388**	0.159
$VFER_{NSR,TKEO}$	**-0.379**	0.309
11^{th} MFCC coef	**0.369**	0.303
$VFER_{NSR,SEO}$	**-0.365**	0.324
4^{th} delta MFCC	**-0.363**	0.219
$VFER_{mean}$	**-0.321**	0.110
RPDE	**0.292**	0.221
DFA	**0.287**	0.324
$Shimmer_{PQ11}$	**0.285**	0.181
HNR_{mean}	**-0.285**	0.315

Ten features most strongly associated with the response, sorted using the magnitude of the correlation coefficient. The correlations are all statistically significant ($p < 0.001$). Also, the results of the Mann Whitney statistical test suggest all relationships are statistically significant ($p < 0.001$). The normalized mutual information lies in the range zero to one, with a value closer to one indicating stronger association. The response was '0' for healthy controls and '1' for people with Parkinson's disease. Thus, positive correlation coefficients suggest that the dysphonia measure takes, in general, larger values for Parkinson's disease phonations.

IV. RESULTS

A. Preliminary Statistical Survey

Table II presents the ten dysphonia features most strongly associated with the response, sorted according to the absolute correlation coefficient value. It is interesting to note that some of the nonlinear dysphonia measures (RPDE, DFA) appear to be quite strongly associated with the response, and exhibit statistically significant ($p < 0.001$) correlation, but the more recently proposed VFER measures, and MFCCs, are more strongly associated. These findings give some initial confidence that the binary classification task of this study has a good chance of success. The statistical correlations between pairs of dysphonia measures (correlation matrix) appear in the online supplementary material.

B. Classification Stage: Mapping Dysphonia Features to the Response

Table III summarizes comparable classification results in the literature, and those in the present study. All the studies cited in Table III used the exact feature data matrix computed in Little *et al*. [5], which comprised 31 subjects (195 phonations) and 22 features. FS was conducted in all of these studies before mapping those (selected) features to the response. Our results are obtained using a larger database with 43 subjects (263 phonations), and a much larger number of features (132) based on the algorithms described in Tsanas *et al*. [11]. For a fair comparison with the original study of Little *et al*. [5], we have also applied the cross-validated classification algorithms of this paper to the optimal feature subset selected in that study.

To date, the best results, across a wide range of classification algorithms, had a reported accuracy of around 93%, when using the same feature data as calculated in [5] (see Table III). Using the 132 features in this study with SVM leads to a noticeable improvement in accuracy (97.7%) over these existing studies. However, these studies used considerably fewer features (at most 22). Therefore, this improved result could be affected by overfitting, and further accuracy gains may occur with fewer features. Thus, we computed the out of sample MAE results using the features selected by the four FS algorithms as the number of features is varied (see Fig. 1). In this way, we found that the globally optimal feature size (minimum MAE) is 22

TABLE III
CLASSIFICATION ACCURACY OF STUDIES IN THE LITERATURE AND THIS PAPER

Study	Learning and validation scheme	Reported accuracy (%)
Guo et al. 2010 [6]	GP-EM, 10-fold cross-validation	93.1 ± 2.9
Das 2010 [35]	Neural network, 35% of the data used for testing following random initial partitioning	92.9
Sakar and Kursun 2010 [36]	SVM, bootstrap with 50 replicates	92.8 ± 1.2
Little et al. 2009 [5]	SVM, bootstrap with 50 replicates	91.4 ± 4.4
Psorakis et al. 2010 [37]	Non-sparse E-M, 10-fold cross-validation with 10 repetitions	89.5 ± 6.6
Shahbaba and Neal 2009 [38]	dpMNL, 5-fold cross-validation	87.7 ± 3.3
*Optimal 4 feature subset from Little et al. 2009 [5]	SVM methodology in this study, 10-fold cross-validation with 100 repetitions, features recalculated	89.3 ± 6.9
*Optimal 4 feature subset from Little et al. 2009 [5]	RF methodology in this study, 10-fold cross-validation with 100 repetitions, features recalculated	89.3 ± 7.2
*All 132 features	SVM, 10-fold cross-validation with 100 repetitions	97.7 ± 2.8
*All 132 features	RF, 10-fold cross-validation with 100 repetitions	90.2 ± 5.9

The results are presented in the form mean ± standard deviation where appropriate. The asterisk (*) indicates new results of the present study. SVM stands for *support vector machine*, dpMNL for *Dirichlet process multinomial logit*, GP-EM for *genetic programming and the expectation maximization algorithm*, E-M for *expectation maximization* algorithm, and RF for *random forests*. All cited studies used the features derived in [5] with 31 subjects; the results in the present study are from an expanded database with 43 subjects, with all features recalculated.

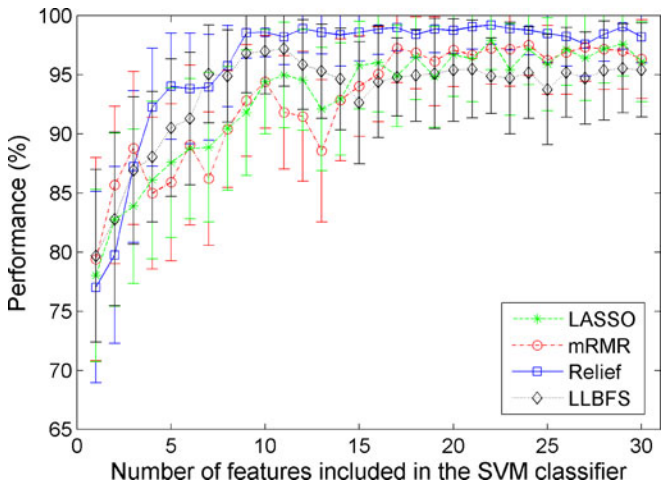

Fig. 1. Comparison of out-of-sample mean performance results with confidence intervals (one standard deviation around the quoted mean performance) using the features selected by each of the four-feature selection algorithms. These results are computed using tenfold CV with 100 repetitions. For clarity, we present here only the first 30 steps.

TABLE IV
SELECTED FEATURE SUBSETS AND CLASSIFICATION PERFORMANCE

LASSO	mRMR	RELIEF	LLBFS
$VFER_{NSR,TKEO}$	2^{nd} MFCC coef	1^{st} MFCC coef	2^{nd} MFCC coef
11^{th} MFCC coef	$Shimmer_{Amplitude,AM}$	11^{th} MFCC coef	11^{th} MFCC coef
$VFER_{NSR,SEO}$	$VFER_{NSR,SEO}$	2^{nd} MFCC coef	9^{th} MFCC coef
4^{th} delta MFCC	$GNE_{NSR,SEO}$	3^{rd} MFCC coef	$VFER_{NSR,TKEO}$
HNR_{mean}	5^{th} delta-delta MFCC	$VFER_{NSR,TKEO}$	$VFER_{entropy}$
GNE_{std}	HNR_{mean}	$VFER_{NSR,SEO}$	$VFER_{NSR,SEO}$
12^{th} MFCC coef	8^{th} MFCC coef	9^{th} MFCC coef	RPDE
RPDE	4^{th} delta MFCC	7^{th} MFCC coef	HNR_{mean}
$OQ_{std\ cycle\ open}$	11^{th} MFCC coef	6^{th} MFCC coef	DFA
2^{nd} MFCC coef	$VFER_{NSR,TKEO}$	8^{th} MFCC coef	4^{th} delta MFCC
94.4 ± 4.4 TP: 97.5 ± 3.4 TN: 86.5 ± 14.3	94.1 ± 3.9 TP: 97.6 ± 3.3 TN: 84.3 ± 13.2	98.6 ± 2.1 TP: 99.2 ± 1.8 TN: 95.1 ± 8.4	97.1 ± 3.7 TP: 99.7 ± 1.7 TN: 89.1 ±13.9

The last row presents the % accuracy when the selected features from each algorithm are fed into the SVM classification algorithm. The results are given in the form mean ± standard deviation and are out of sample computed using tenfold cross validation with 100 repetitions. TP stands for true positive (PWP) and TN for true negative (healthy controls).

using RELIEF, but this is not a practically useful improvement over the MAE when using only ten features. Following the principle of parsimony then, we choose the least number of features giving the most accurate results according to mean performance (%). Therefore, our subsequent results use only the first ten features (see Table IV) for each FS algorithm (the features are presented in descending order of selection).

The SVM also outperforms RF in this reduced feature space (for example, using the ten features from RELIEF in Table IV, RF achieves only 93.5% accuracy compared to 98.6% accuracy with SVM). We remark that reducing the original 132-dimensional feature space can lead to an *improvement* in out-of-sample performance accuracy with both SVM and RF. Overall, these findings suggest that we can estimate whether someone has PD or is healthy from a single phonation, with almost 99% accuracy using only ten dysphonia features, a considerable improvement over previous results.

Finally, we examine whether the out-of-sample results using different FS algorithms (see Table IV) are statistically significantly different. Specifically, we compared the distributions of the classification errors obtained using RELIEF against the distributions of classification errors with the alternative FS approaches (Mann–Whitney rank sum test). In all three cases, the test rejected the null hypothesis of equal medians ($p < 0.001$); hence, the classification results using RELIEF-selected features are statistically significantly better from the results obtained using the other FS algorithms.

V. DISCUSSION

Decision support tools in biomedical applications are generating considerable research interest not least because of their potential to improve healthcare provision. In this study, we have applied an extensive range of classical and novel speech signal processing algorithms for vocal pathology assessment in order to investigate how to discriminate PWP from healthy controls using sustained vowel phonations. This binary discrimination problem has attracted interest in recent years, with the best results reporting approximately 93% classification accuracy on a subset of 22 features. Here, we demonstrated that we can achieve almost 99% accuracy using ten dysphonia measures. Compared to previous studies in this application, we have used an expanded speech database (which included all the 195 phonations in the original database and 68 additional phonations), and introduced many recently proposed dysphonia measures, which have not been previously used in this application (all the dysphonia measures in this study were computed anew using the algorithms described in [11]). As in previous studies, we have used nonlinear SVMs for mapping features to the response, and also investigated RF.

A novel contribution in this paper is to use four different FS algorithms to find a small subset of only ten features from the original 132. This led to an informative feature subset for the binary classification task of this study, which may also tentatively suggest the most detectable characteristics of voice impairment in PD. All FS algorithms coped relatively well with the task, but RELIEF provided the subset with the lowest classification error. Recent research has demonstrated that RELIEF may work very well, in practice, in this kind of application because, internally, it incorporates a (nonlinear, nearest-neighbor) classifier [30]. The presence of highly correlated features (see the Excel file in the online supplementary material) indicates that LASSO may not be in its optimal setting (sparse environment with low feature correlations) to perform well. Thus, LASSO may be selecting some noisy features, which may not assist the discrimination of the two classes. Recently, we have found that feature complementarity may be a required aspect of FS in a related application [31]. Therefore, mRMR, which does not take into account feature complementarity, may also not be the most appropriate algorithm in this application. These insights may help explain why RELIEF and LLBFS appear to work better in this domain.

One interesting new finding is that of all the families of measures tested here, MFCCs and SNR measures (VFER, HNR, GNE) appear to be consistently selected (see Table IV). The pathophysiological importance of SNR measures is well known: it is most likely the effect of amplified aeroacoustic noise due to increased airflow turbulence, ultimately generated by incomplete vocal fold closure. However, the selection of MFCCs is somewhat surprising, because these measures are mainly sensitive to insufficient control in the steady placement of the articulators, which amplify specific acoustic resonances and attenuate others in the vocal tract. This may indicate that more research into the effect of PD on vocal tract articulatory impairment, even for sustained phonations, is required. By design, MFCCs are not highly correlated (see the correlation matrix in the online

supplementary material), and provide *complementary* information regarding characteristics of the speech signal. Combined with the fact that some MFCCs are relatively highly correlated with the response (see Table II), provides a highly plausible explanation for why RELIEF tends to select these features. Compared to the original study of Little *et al*. [5] where the selected feature subset comprised HNR, RPDE, DFA, and PPE, RELIEF consistently selected the new dysphonia measures presented here. LLBFS (the FS algorithm which resulted in the second best performance) selected RPDE, HNR, and DFA with lower rank (7–9) compared to the new features described here. These findings justify the higher classification accuracy obtained in this study in comparison to previous studies.

In our experiments, SVM has a clear edge over RF for this particular application (see Table III). We also verified Breiman's observation [32] that modifying the RF tuning parameter (the number of features over which to search to construct each branch of each tree) does not produce markedly different results in the overall RF classification accuracy. Some empirical studies have compared SVM and RF with no clear verdict about overall superiority of either approach [39], although it is well established that both classifiers perform well in general [22]. It would be interesting to investigate the reasons that RF perform noticeably worse than SVM in this application. As Statnikov *et al*. [40] remark, this undertaking is not straightforward, and requires extensive empirical and theoretical studies to explain the performance differences observed across different studies for SVMs and RF [36]. Moreover, it may be worth taking into account the confidence of the classifiers' decisions. Both SVMs and RF can be arranged to produce probabilistic outputs, and it would be possible to introduce an additional "Don't know" class if the probability of the class assignments was below some prespecified threshold. In a practical setting, assigning probabilities to an automatic decision support tool would aid clinicians in deciding upon further actions.

It has recently been suggested that it may be useful to partition the data according to gender in a similar application (mapping the dysphonia measures to a clinical metric that quantifies average Parkinson's disease symptom severity [11]). Here, this would require an entirely different dysphonia feature subset and classifier for males versus females. However, reducing the available data by splitting the original dataset into two subsets diminishes the *statistical power* of the performance evaluations. When we attempted data partitioning according to gender with this data, we obtained reduced performance accuracy. We emphasize that with more data, it is possible that partitioning (which may or may not be limited to gender partitioning) could lead to interesting insights. For example, data partitioning by gender could provide insight into the most useful features for males versus females with regard to the discrimination of PWP from healthy controls, as in Tsanas *et al*. [11].

We envisage this study as a step toward the larger goal of technologies for diagnostic decision support in PD. The algorithms in this study appear to be very effective for discriminating PWP from healthy controls on the basis of extensive CV tests. Conceptually, CV provides an estimate of the performance of the model on new data, assuming that the new dataset is drawn from the same distribution as the dataset used to train the classifier. Therefore, the findings of this study might need to be further validated using independent datasets before this technology could be used as a diagnostic decision support tool. We are working toward collecting new datasets toward this aim. Furthermore, we remark that the healthy subjects in this study did not have any pathological vocal symptoms when assessed by expert speech scientists. A study involving a cohort of subjects with PD-like vocal symptoms, but without PD, would further validate the applicability of these findings. Although running speech has been used in other studies [7], the collection of sustained vowels in controlled circumstances reduces intraspeaker variability and confounding linguistic factors, and may lead to better results. Nevertheless, future studies could investigate the combination of both approaches, extracting information from both sustained vowels and running speech. It would be interesting to use a very large database including voices from diverse disorders, where the use of sophisticated dysphonia measures might help determine the underlying pathology amongst a wide set of possible diagnoses. Also, the data in this study are collected in an acoustically controlled environment; we are currently working to extend these findings to more realistic acoustic setups, which would extend the proposed technology for use in more practical settings. Finally, future work could incorporate additional information from physical models of voice production mechanisms, for example to improve the accuracy of jitter, shimmer and HNR estimates using glottal source signals obtained from the voice recordings.

ACKNOWLEDGMENT

The first author would like to thank Dr. Y. Sun for providing the source code for the LLBFS method.

REFERENCES

[1] M. Rajput, A. Rajput, and A. H. Rajput, "Epidemiology," in *Handbook of Parkinson's disease*, R. Pahwa and K. E. Lyons, Eds., 4th ed. New York: Informa Healthcare, 2007, ch. 2.

[2] A. S. von Campenhausen, B. Bornschein, R. Wick, K. Bötzel, C. Sampaio, W. Poewe, W. Oertel, U. Siebert, K. Berger, and R. Dodel, "Prevalence and incidence of Parkinson's disease in Europe," *Eur. Neuropsychopharmacol*., vol. 15, pp. 473–490, 2005.

[3] A. Schrag, Y. Ben-Schlomo, and N. Quinn, "How valid is the clinical diagnosis of Parkinson's disease in the community?," *J. Neurol., Neurosurg. Psychiat*., vol. 73, pp. 529–535, 2002.

[4] S. K. Van Den Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter, D. A. Bloch, and L. Nelson, "Incidence of Parkinson's disease: Variation by age, gender, and race/ethnicity," *Am. J. Epidem*., vol. 157, pp. 1015–1022, 2003.

[5] 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.

[6] P.-F. Guo, P. Bhattacharya, and N. Kharma, "Advances in detecting Parkinson's disease," *Med. Biometrics (Lect. Not. Comput. Sci.)*, vol. 6165, pp. 306–314, 2010.

[7] S. Sapir, L. Ramig, J. Spielman, and C. Fox, "Formant Centralization Ratio (FCR): A proposal for a new acoustic measure of dysarthric speech," *J. Speech Language Hearing Res*., vol. 53, pp. 114–125, 2010.

[8] A. Ho, R. Iansek, C. Marigliani, J. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behav. Neurol*., vol. 11, pp. 131–137, 1998.

[9] 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 Cognition*, vol. 56, pp. 24–29, 2004.

[10] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression using non-invasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, Apr. 2012.

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

[12] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, 2010, pp. 594–597.

[13] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity," in *Proc. Int. Symp. Nonlinear Theory Appl.*, Krakow, Poland, Sep. 5–8, 2010, pp. 457–460.

[14] R. J. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. San Diego, CA: Singular Thomson Learning, 2000.

[15] I. R. Titze, *Principles of Voice Production*, 2nd ed. Iowa City: Natl. Center Voice Speech, 2000.

[16] J. Schoentgen and R. De Gucteneere, "Time series analysis of jitter," *J. Phonetics*, vol. 23, pp. 189–201, 1995.

[17] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[18] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for non-linear and non stationary time series analysis," *Proc. Roy. Soc. London A*, vol. 454, pp. 903–995, 1998.

[19] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct. 2006.

[20] R. Fraile, N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz, and C. Fredouille, "Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex," *Folia Phoniatrica et Logopaedica*, vol. 61, pp. 146–152, 2009.

[21] A. Ihler and M. Mandel, *KDE Toolbox,* (2003). [Online]. Available: http://www.ics.uci.edu/∼ihler/code/kde.html.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[23] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh, and Eds, *Feature Extraction: Foundations and Applications*. New York: Springer, 2006.

[24] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. B*, vol. 58, pp. 267–288, 1996.

[25] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[26] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. 9th Int. Conf. Mach. Learn.*, 1992, pp. 249–256.

[27] Y. Sun, S. Todorovic, and S. Goodison, "Local learning based feature selection for high dimensional data analysis," *IEEE Pattern Anal. Mach.*, vol. 32, no. 9, pp. 1610–1626, Sep. 2010.

[28] D. Donoho, "For most large underdetermined systems of equations, the minimal L1-norm near-solution approximates the sparsest near-solution," *Commun. Pure Appl. Math.*, vol. 59, no. 7, pp. 904–934, 2006.

[29] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Ann. Statist.*, vol. 37, no. 1, pp. 246–270, 2009.

[30] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection—Theory and algorithms," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, pp. 43–50.

[31] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Robust parsimonious selection of dysphonia measures for telemonitoring of Parkinson's disease symptom severity," in *Proc. 7th Int. Workshop Models Anal. Vocal Emissions Biomed. Appl. (MAVEBA)*, Florence, Italy, Aug. 25–27, 2011, p. 169—172.

[32] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, p. 5-32, 2001.

[33] C. C. Chang and C-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, 2011.

[34] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," National Taiwan University, Taipei, Taiwan, Tech. Rep., 2010.

[35] R. Das, "Classification of Parkinson's disease by using voice measurements," *Expert Syst. Appl.*, vol. 37, pp. 1568–1572, 2010.

[36] C. O. Sakar and O. Kursun, "Telediagnosis of Parkinson's disease using measurements of dysphonia," *J. Med. Syst.*, vol. 34, pp. 591–599, 2010.

[37] I. Psorakis, T. Damoulas, and M. A. Girolami, "Multiclass relevance vector machines: Sparsity and accuracy," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1588–1598, Oct. 2010.

[38] B. Shahbaba and R. Neal, "Nonlinear models using Dirichlet process mixtures," *J. Mach. Learn. Res.*, vol. 10, pp. 1829–1850, 2009.

[39] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, pp. 169–186, 2003.

[40] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC Bioinformatics*, vol. 9, p. 319, 2008.

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