Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease

Max A. Little*, *Member, IEEE*, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig

***Abstract*—In this paper, we present an assessment of the practical value of existing traditional and nonstandard measures for discriminating healthy people from people with Parkinson's disease (PD) by detecting dysphonia. We introduce a new measure of dysphonia, pitch period entropy (PPE), which is robust to many uncontrollable confounding effects including noisy acoustic environments and normal, healthy variations in voice frequency. We collected sustained phonations from 31 people, 23 with PD. We then selected ten highly uncorrelated measures, and an exhaustive search of all possible combinations of these measures finds four that in combination lead to overall correct classification performance of 91.4%, using a kernel support vector machine. In conclusion, we find that nonstandard methods in combination with traditional harmonics-to-noise ratios are best able to separate healthy from PD subjects. The selected nonstandard methods are robust to many uncontrollable variations in acoustic environment and individual subjects, and are thus well suited to telemonitoring applications.**

***Index Terms*—Biomedical measurements, nervous system, speech analysis, telemedicine.**

I. Introduction

NEUROLOGICAL disorders, including Parkinson's disease (PD), Alzheimer's, and epilepsy, profoundly affect the lives of patients and their families. PD affects over one million people in North America alone [3]. Moreover, an aging population means this number is expected to rise as studies suggest rapidly increasing prevalence rates after the age of 60 [4]. In addition to increased social isolation, the financial burden of PD is significant and is estimated to rise in the future [5]. Currently, there is no cure, although medication is available offering significant alleviation of symptoms, especially at the early stages of the disease [6]. Most *people with Parkinson's* (PWP) disease will therefore be substantially dependent on clinical intervention.

For many PWP, the requisite physical visits to the clinic for monitoring and treatment are difficult. Widening access to the Internet and improved telecommunication systems bandwidth offer the possibility of remote monitoring of patients (*telemedicine* [7]), with substantial opportunities for lowering the inconvenience and cost of physical visits. However, in order to exploit these opportunities, there is the need for reliable clinical monitoring tools. Research has shown that approximately 90% of PWP exhibit some form of vocal impairment [9], [10]. Vocal impairment may also be one of the earliest indicators for the onset of the illness [11], and the measurement of voice is noninvasive and simple to administer. Thus, voice measurement to detect and track the progression of symptoms of PD has drawn significant attention [13], [14].

PWP typically display a constellation of vocal symptoms that include impairment in the normal production of vocal sounds (*dysphonia*) and problems with the normal articulation of speech (*dysarthria*) (see [15] and references therein for a comprehensive description of these symptoms). Dysphonic symptoms typically include reduced loudness, breathiness, roughness, decreased energy in the higher parts of the harmonic spectrum, and exaggerated vocal tremor.

There are many vocal tests that have been devised to assess the extent of these symptoms. These include *sustained phonations* [16], [17], where the patient is instructed to produce a single vowel and hold the pitch of this as constant as possible, for as long as possible, and *running speech* tests [17] where patients are instructed to speak a standard sentence constructed to contain a representative sample of linguistic units. Several of these tests may need to be administered for a full assessment of vocal impairment, but any symptom is sufficient for detecting the severity of PD. Although running speech might be considered a more realistic test of impairment in actual everyday usage, simple sustained phonation tests are able to elicit dysphonic symptoms, and tests of the effectiveness of measurements for detecting dysphonia are best conducted without the confounding effects of articulatory or linguistic components of running speech. Therefore, in this study, we will concentrate on sustained phonation tests.

There have been extensive studies of speech measurement for general voice disorders [8], [18]–[23] and PD in particular [14], [24]. Speech sounds produced during standard speech tests are recorded using a microphone, and the recorded speech signals are subsequently analyzed using measurement methods (implemented in software algorithms) designed to detect certain properties of these signals. The main traditional measurement methods include F0 (the fundamental frequency or *pitch* of vocal oscillation), absolute sound pressure level (indicating the relative loudness of speech), *jitter* (the extent of variation in speech F0 from vocal cycle to vocal cycle), *shimmer* (the extent of variation in speech amplitude from cycle to cycle),

Manuscript received April 25, 2008; revised July 12, 2008. First published September 30, 2008; current version published May 6, 2009. This work was supported by the National Institute of Health (NIH) under Grant NIH-NIDCD R01-DC1150. *Asterisk indicates corresponding author*.

*M. A. Little is with the Systems Analysis, Modeling and Prediction Group, University of Oxford, Oxford, OX1 2JD U.K. (e-mail: littlem@ieee.org).

P. E. McSharry is with the Systems Analysis, Modeling and Prediction Group, University of Oxford, Oxford, OX1 2JD U.K.

E. J. Hunter and J. Spielman are with the National Center for Voice and Speech, Denver Center for the Performing Arts, Denver, CO 80204 USA.

L. O. Ramig is with the National Center for Voice and Speech, The Denver Center for the Performing Arts, Denver, CO 80204 USA, and also with the Department of Speech, Language and Hearing Sciences, University of Colorado at Boulder, Boulder, CO 80309 USA.

Digital Object Identifier 10.1109/TBME.2008.2005954

and *noise-to-harmonics ratios* (the amplitude of noise relative to tonal components in the speech) [16]. Studies have shown variations in all these measurements for PWP by comparison to healthy controls [25], indicating that these could be useful measures in assessing the extent of dysphonia.

More recently, a variety of novel measurement methods have been devised to assess dysphonic symptoms, in particular, those based on *nonlinear dynamical systems* theory [12], [26]. These measurements are motivated by extensive modeling studies [27] and evidence [28] that vocal production is a highly nonlinear dynamical system, and that changes caused by impairments to the vocal organs, muscles, and nerves will affect the dynamics of the whole system. As a result, these changes can be detected by *nonlinear time series analysis* tools [12], such as *correlation dimension* and methods for characterizing *pseudoperiodic* time series [29], [30]. Similarly, randomness and noise are inherent to vocal production [8]; as a result, tools such as *recurrence period density entropy* (RPDE) and *detrended fluctuation analysis* (DFA) have been applied to speech signals, showing the ability to detect general voice disorders [8].

Nonetheless, practical, remote assessment of dysphonia requires high reliability, and this is impeded by several confounding issues. Sound recording and measurement methods will differ in robustness to uncontrolled variation in the acoustic environment of the clinic and home and to the physical condition and characteristics of the subject. In order to gain as much reliability as possible, measurement methods should be chosen that are as robust as possible to such uncontrolled (and in many cases, uncontrollable) variations. For example, absolute sound pressure-level measurement requires costly calibration equipment, and the requisite precision is often difficult to obtain. This limits the reliability of this measure in telemedicine applications. Similarly, although PD-related dysphonia is associated with reduced absolute speech F0, this is confounded by unrelated effects such as individual preferences or subject gender [24].

Although there are a large number of traditional and novel measurement methods for the assessment of voice disorders and the character of PD-specific dysphonia is fairly well established, there are no methods for efficiently characterizing such dysphonia in the presence of known confounding factors such as subject gender and highly variable acoustic environments. For this reason, we introduce a new measure of dysphonia that we dub *pitch period entropy* (PPE), a robust measure sensitive to observed changes in speech specific to PD. Statistically significant relationships have been shown to exist between the extent of dysphonia in PD and measurement methods [14]. Nonetheless, in remote monitoring conditions, we can expect much more variation in these measurements than the controlled conditions under which these studies were conducted. Given the need for high reliability in telemedicine applications therefore, we must assess the *practical relevance* of the variation in measurements with severity of dysphonia in PD. Statistical significance alone is not sufficient, as this does not give a complete picture of the extent to which any one measurement or set of measurements is useful in determining the extent of PD-related dysphonia [31].

Methods from statistical learning theory, such as *linear discriminant analysis* (LDA) and *support vector machines* (SVMs) [32], are preferred here because they can directly measure the extent to which PWP can be discriminated from healthy controls on the basis of measures of dysphonia, addressing the problem of *classifying* subjects as healthy or PD. With such classification methods, it is also possible to combine measures to create more effective discrimination in practice. Measures from each subject are placed together in a (multidimensional) *feature vector* that forms the input to the classification method [32]. The method finds a *decision boundary* in the *feature space* formed by these vectors, so that the class of each subject (healthy or PD) can be predicted on the basis of subsequent voice measures. The rate of correct classification can be used to assess which measures contain the most useful information to best separate healthy from PWP in remote monitoring applications. This also allows us to assess the value of traditional with novel nonlinear and/or stochastic methods of dysphonia measurement for PD [33].

Nonetheless, given the very large number of measures of dysphonia, it is computationally infeasible to test all possible combinations. Furthermore, theoretical considerations show that as the feature set size increases, reliable classification is impaired by the diminished coverage of the feature space with measures from a fixed number of subjects [32]. Some form of *feature selection* must therefore be practiced [34] to reduce the set of measures down to a minimal size that contains the optimal amount of information for effective classification. Unfortunately, nothing short of a full, exhaustive (but intractable) search is guaranteed to produce the optimal feature set [34]. As a compromise, in this study, we first apply a preselection *filter* that removes redundant measures, followed by an exhaustive search, testing all possible combinations of the filtered measures with an SVM classifier.

The paper is organized as follows. The speech data used in this study is described in Section II, and the various methods of speech measurement, preprocessing, preselection, and classification are presented in Section III. In Section IV, we present the results of our findings in comparing the various techniques. Section V discusses the interpretation of these findings, and provides conclusions and relevance of the results for future telemedicine applications.

II. Data

The data for this study consist of 195 sustained vowel phonations from 31 male and female subjects, of which 23 were diagnosed with PD. The time since diagnoses ranged from 0 to 28 years, and the ages of the subjects ranged from 46 to 85 years (mean 65.8, standard deviation 9.8). Averages of six phonations were recorded from each subject, ranging from 1 to 36 s in length. See Table I for subject details. Fig. 1 shows plots of two of these speech signals. The phonations were recorded in an Industrial Acoustics Company (IAC) sound-treated booth using a head-mounted microphone (AKG C420) positioned at 8 cm from the lips. The microphone was calibrated as described in [35] using a class 1 sound-level meter (B&K 2238) placed 30 cm from the speaker. The voice signals were recorded directly on computer using Computerized Speech Laboratory (CSL) 4300B hardware (Kay Elemetrics), sampled at 44.1 kHz, with

TABLE I
LIST OF SUBJECTS WITH SEX, AGE, PARKINSON'S STAGE, AND NUMBER OF YEARS SINCE DIAGNOSIS

Subject code	Sex	Age	Stage (H&Y)	Years since diagnosis
S01	M	78	3.0	0
S34	F	79	2.5	¼
S44	M	67	1.5	1
S20	M	70	3.0	1
S24	M	73	2.5	1
S26	F	53	2.0	1½
S08	F	48	2.0	2
S39	M	64	2.0	2
S33	M	68	2.0	3
S32	M	50	1.0	4
S02	M	60	2.0	4
S22	M	60	1.5	4½
S37	M	76	1.0	5
S21	F	81	1.5	5
S04	M	70	2.5	5½
S19	M	73	1.0	7
S35	F	85	4.0	7
S05	F	72	3.0	8
S18	M	61	2.5	11
S16	M	62	2.5	14
S27	M	72	2.5	15
S25	M	74	3.0	23
S06	F	63	2.5	28
S10 (healthy)	F	46	n/a	n/a
S07 (healthy)	F	48	n/a	n/a
S13 (healthy)	M	61	n/a	n/a
S43 (healthy)	M	62	n/a	n/a
S17 (healthy)	F	64	n/a	n/a
S42 (healthy)	F	66	n/a	n/a
S50 (healthy)	F	66	n/a	n/a
S49 (healthy)	M	69	n/a	n/a

Entries labeled "n/a" for healthy subjects for which Parkinson's stage and years since diagnosis is not applicable. "H&Y" refers to the Hoehn and Yahr PD stage, where higher values indicate greater level of disability [2].

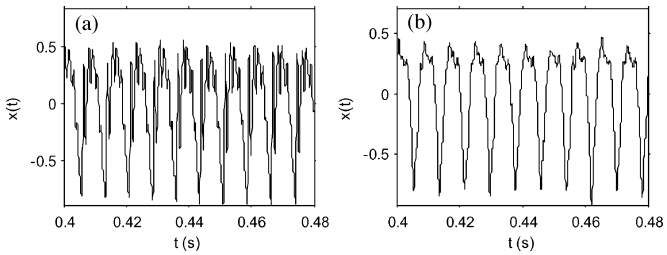

Fig. 1. Two selected examples of speech signals. (a) Healthy. (b) Subject with PD. The horizontal axis is time in seconds and the vertical axis is signal amplitude (no units).

16-bit resolution. Although amplitude normalization affects the calibration of the samples, the study is focused on measures insensitive to changes in absolute speech pressure level. Thus, to ensure robustness of the algorithms, all samples were digitally normalized in amplitude prior to calculation of the measures.

III. METHODS

As discussed in Section I, the methodology of this study can be broken down into three stages: 1) the calculation of features; 2) the preprocessing and preselection of features; and (c) the application of a classification technique to all possible subsets of features for the discrimination of healthy from disordered subjects, selecting the subset that produces the best classification performance.

TABLE II
LIST OF MEASUREMENT METHODS APPLIED TO ACOUSTIC SIGNALS RECORDED FROM EACH SUBJECT

Feature	Retained after filtering?	Description
MDVP:Jitter(%)	No	Kay Pentax MDVP jitter as a percentage [1]
MDVP:Jitter(Abs)	Yes	Kay Pentax MDVP absolute jitter in microseconds [1]
MDVP:RAP	No	Kay Pentax MDVP Relative Amplitude Perturbation [1]
MDVP:PPQ	No	Kay Pentax MDVP five-point Period Perturbation Quotient [1]
Jitter:DDP	Yes	Average absolute difference of differences between cycles, divided by the average period [1]
MDVP:Shimmer	No	Kay Pentax MDVP local shimmer [1]
MDVP:Shimmer(dB)	No	Kay Pentax MDVP local shimmer in decibels [1]
Shimmer:APQ3	No	Three point Amplitude Perturbation Quotient [1]
Shimmer:APQ5	No	Five point Amplitude Perturbation Quotient [1]
MDVP:APQ	Yes	Kay Pentax MDVP 11-point Amplitude Perturbation Quotient [1]
Shimmer:DDA	Yes	Average absolute difference between consecutive differences between the amplitudes of consecutive periods [1]
NHR	Yes	Noise-to-Harmonics Ratio [1]
HNR	Yes	Harmonics-to-Noise Ratio [1]
RPDE	Yes	Recurrence Period Density Entropy [8]
DFA	Yes	Detrended Fluctuation Analysis [8]
D2	Yes	Correlation dimension [12]
PPE	Yes	Pitch period entropy [this paper]

MDVP stands for (Kay Pentax) Multidimensional voice program. See main text for detailed descriptions of the algorithms used to calculate these features.

A. Feature Calculation Stage

The feature calculation stage involves the application of a representative selection of traditional and nonstandard measurement methods to all the speech signals. Each method produces a single number for each of the 195 signals. See Table II for a list of the measures used as features in this study.

1) Calculation of Traditional Measures: Calculation of the traditional measures was performed using the software Praat [1]. To facilitate comparison with other studies, where possible, traditional measures were chosen that coincide with an equivalent measure computed by the Kay Pentax multidimensional voice program (MDVP) [36]. These measures are prefixed "MDVP." The traditional measures are based on the application of the short-time autocorrelation to successive segments of the signal, with peak picking to determine the frequency of vibration of the vocal folds (F0 or *pitch period*), and location in time of the beginning of each cycle of vibration of the vocal folds (*pitch marks*) [37].

The jitter and period perturbation measures are derived from the sequence of frequencies for each vocal cycle, by taking successive absolute differences between frequencies of each cycle and averaging over a varying number of cycles, optionally normalizing by the overall average. The shimmer and

amplitude perturbation measures are derived from the sequence of maximum extent of the amplitude of the signal within each vocal cycle. The average difference of this sequence is taken as a measure of the deviation between cycle amplitudes. The noise-to-harmonics (and harmonics-to-noise) ratios are derived from the signal-to-noise estimates from the autocorrelation of each cycle. See [1], [36], and [37] for more details of the calculation of these traditional measures.

In order to increase the power of these algorithms in separating healthy from PWP, we discard the second half of each voice signal in calculating these measures. This is because the end of the phonation is dominated by spurious dysphonia caused mainly by lack of lung pressure. Many PWP exhibit similar dysphonia, which otherwise would be conflated with dysphonia caused by natural lack of lung pressure. Although other studies have found statistical relationships between absolute values of F0 and PD-related dysphonia, we do not use this as a measure because it is adversely affected by gender and individual differences. Similarly, although it is observed that lower absolute sound pressure levels (amplitudes) are associated with PD-related dysphonia, for practical reasons, we do not use this as a measure because the precision calibration required to obtain reliable estimates of this quantity are difficult to achieve in remote monitoring situations. Thus, here we are deliberately restricted to relative (or *perturbative*) measures of pitch period and amplitude since they are more robust to uncontrollable environmental and individual variations.

2) Calculation of Nonstandard Measures: The correlation dimension (D2) is calculated by first *time-delay embedding* the signal to recreate the phase space of the nonlinear dynamical system that is proposed to generate the speech signal [12]. In this reconstructed phase space, a geometrically *self-similar* (*fractal*) object indicates complex dynamics, which are implicated in dysphonia [38]. We use the Time SEries ANalysis (TISEAN) implementation [39]. The *recurrence period density entropy* (RPDE) quantifies the extent to which dynamics in the reconstructed phase space after time-delay embedding can be considered as strictly periodic, i.e., repeating exactly [8]. A *recurrent* signal returns to the same point in the phase space after a certain length of time, called the *recurrence period T*. It has been shown that the deviation from periodicity evaluated by the *entropy H* of the distribution of these recurrence periods $P(T)$ is a good indicator of general voice disorders, as general voice pathologies lead to impairment in the ability to sustain regular vibration of the vocal folds [8]. Dividing through by the entropy of the uniform distribution normalizes the RPDE values (H_{norm}) to the range [0, 1].

Finally, DFA is a measure of the extent of the *stochastic self-similarity* of the noise in the speech signal. The noise in speech is mostly generated by turbulent airflow through the vocal folds [40]. Such turbulent processes are characterized by a statistical scaling exponent α on a range of physical scales, which manifests in measured aspects of the dynamics including acoustic pressure fields. In some voice disorders, incomplete vocal fold closure leads to changes in this turbulent "breath" noise, and the characteristics of the self-similarity of the noise in the speech signal is therefore an indicator of dysphonia [8]. It is found that for general voice disorders, the scaling exponent is larger for dysphonic than healthy subjects [8], [19]. The DFA algorithm calculates the extent of amplitude variation $F(L)$ of the speech signal over a range of time scales L, and the self-similarity of the speech signal is quantified by the slope α of a straight line on a log–log plot of L versus $F(L)$. A simple nonlinear transformation then normalizes these slope values (α_{norm}) to the range [0, 1] [8].

3) New Measure of PD Dysphonia (PPE): All healthy voices exhibit natural pitch (F0) variation characterized by smooth *vibrato* and *microtremor* [41], and this is detected in traditional jitter measures, for example. However, one common dysphonic PD symptom is impaired control of stationary voice pitch (F0) during sustained phonation [24]. Thus, with traditional measures, it is difficult to separate natural, healthy pitch variations from dysphonic variations due to PD. Similarly, the extent of this natural variation is related to the average voice pitch of the subject; speakers with naturally high-pitched voices will have much larger vibrato and microtremor than those with low-pitched voices, when these variations are measured on an absolute frequency (in hertz) scale. Therefore, measurements of abnormal speech pitch variation need to take into account these two important effects: healthy, smooth vibrato and microtremor, and the *logarithmic* nature of speech pitch in speech production (and perception). These observations suggest that a more relevant scale on which to assess abnormal variations in speech pitch is the perceptually relevant, *logarithmic* (*tonal*) scale, rather than the absolute frequency scale [42]. It also suggests that in order to better capture pitch period variation due to PD-related dysphonia independent of these natural variations, smooth variations should be removed prior to measuring the extent of such variations.

To implement these two insights algorithmically, we first obtain the pitch sequence of the phonation and convert to the logarithmic *semitone* scale $p(t)$, where p is the semitone pitch at time t. We next analyze the *roughness* of variations in this sequence over and above any healthy, smooth variations, by first removing linear temporal correlations in this semitone sequence with a standard linear whitening filter (coefficients of which are estimated using linear prediction by the covariance method [43]) to produce the relative semitone variation sequence $r(t)$. This filtering effectively flattens the spectrum of the semitone time series and removes the effect of the mean semitone (which depends on the individual preferences and gender). Subsequently, we construct a discrete probability distribution of occurrence of relative semitone variations $P(r)$. Finally, we calculate the *entropy* of this probability distribution [44], which then characterizes the extent of (non-Gaussian) fluctuations in the sequence of relative semitone pitch period variations. An increase in this entropy measure better reflects the variations over and above natural healthy variations in pitch observed in healthy speech production.

B. Feature Preparation and Classification Stage

Practical exploitation of the information in the measures calculated before requires us to construct feature vectors from these

measures, which can then be subsequently used to discriminate healthy from PWP. SVM classification performance is greatly enhanced by preprocessing of the values of each measure with an appropriate rescaling [32]. Here, we scale each measure such that, over all signals, the measure occupies the numerical range $[-1, 1]$.

Also, in this stage, we wish to filter the number of measures down to a manageable size, such that a full search of all possible combinations can be conducted [34] in order to determine the optimal set for classification. We note that many of the measures will be highly correlated with other measures. This is because they will be measuring very similar aspects of the speech signal, for example, Jitter(%) and Jitter(Abs) (see Table I) are derived from pitch period sequences and measure the average absolute temporal differences in these periods. Because of this correlation, only one of this pair of measures will contribute useful information for the classification stage and the other should be removed. We therefore systematically search through all pairs of features. Of those that are highly correlated (with a correlation coefficient of greater than 0.95), we remove one of the pairs. We then construct feature vectors with each possible combination of subsets of preprocessed, filtered measures. To each combination, we apply SVM classification. This is a direct measure of the practical separability of the classes.

Prior visual inspection of the layout and clustering of pairs of measures indicate that the optimal decision boundaries separating healthy from PWP may not be simple lines or hyperplanes. Because of this, we use the kernel-SVM formulation, with *Gaussian radial basis kernel* functions [32]. These are flexible kernels that allow smooth, curved decision boundaries. For each combination of features, the classification performance is assessed in terms of the overall number of subjects correctly classified as healthy or PD, the number of PWP correctly classified (the *true positive rate*), and the number of healthy subjects correctly classified (the *true negative rate*). Validation of the results to obtain an estimate of out-of-sample performance and confidence intervals is assessed using *bootstrap resampling* with 50 replicates [32]. The choice of optimal SVM penalty value and kernel bandwidth is determined by exhaustive search over a range of values. The bootstrap classification produces a set of classification performance results for each bootstrap replicate. In order to determine the best performing subset of features, we compare the sets of overall classification results using the two-sided Wilcoxon rank-sum test against the null hypothesis of equal medians, at a significance probability of 0.05.

IV. Results

A. Feature Calculation

There is considerable variation in the distribution of values of the measures. Most of the traditional jitter and shimmer measures produce values close to zero, whereas the novel, nonstandard measures and harmonics-to-noise ratios are more evenly spread over a wider range of values. Fig. 2 shows the results of calculating the RPDE and DFA values for some selected speech signals. As can be seen, for healthy subjects, the recurrence period density $P(T)$ shows a single peak near the time T at which the voice signal tends to repeat itself. For many PWP, however, the recurrence periods are spread over a wide range of values, which indicates that the vocal folds are not oscillating at regular intervals. This is likely caused by impairment of the stable positioning of the intrinsic laryngeal muscles (those that directly move the vocal folds), or extrinsic laryngeal muscles (connecting the larynx and other structures), or by weakness in the production of stable airflow from the lungs.

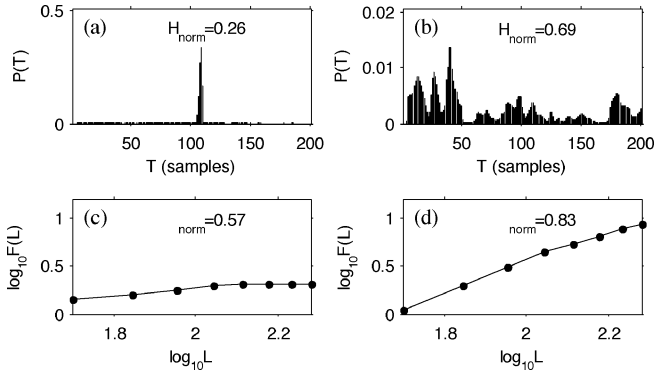

Fig. 2. RPDE and DFA results for healthy subjects (left panels) and for subjects with Parkinson's (right panels). (a) and (b) Recurrence period density P(T) for recurrence times T. (c) and (d) Log–log plot of scaling window sizes L against fluctuation amplitudes F(L). See main text for more detailed descriptions.

For many healthy subjects, the energy in the airflow of the lungs is well imparted to the movement of the vocal folds to generate clear sustained phonations. Thus, the speech signal will be smoother, and this is shown in the smaller DFA scaling exponent. However, many PWP are unable to maintain stable vocal fold vibration, and much more of the airflow energy will be transferred to turbulent acoustic noise generation mechanisms. Hence, the speech signal will be rougher, and this can be seen in an increase in the DFA scaling exponent. Regarding the PPE measure (in Fig. 3), we can see that healthy semitone pitch sequences tend to be quite stable with signs of small, regular, smooth vibrato, and microtremor. After removing this healthy variation with the whitening filter, the distribution of residuals shows a strong peak at zero. The entropy of this distribution is correspondingly small. For PWP, however, the semitone pitch sequence shows considerable irregular variation; the whitened sequence is extremely rough and the distribution of residuals is spread over a wide range of values. This is picked up by the large entropy value.

B. Feature Preparation and Classification

After preprocessing by range scaling, Fig. 4 shows distributions estimated using the Gaussian kernel density method for a representative selection of the measures.

The jitter and shimmer measure values are all very close to zero, with some rare examples of exceptionally high values. The other measures are more evenly spread over the full range of values. The nonstandard measures show more distinction between the mode of the values for healthy controls and PWP, whereas the modes of the harmonics-to-noise ratio values are not as well separated. Fig. 5 shows that some of the measures

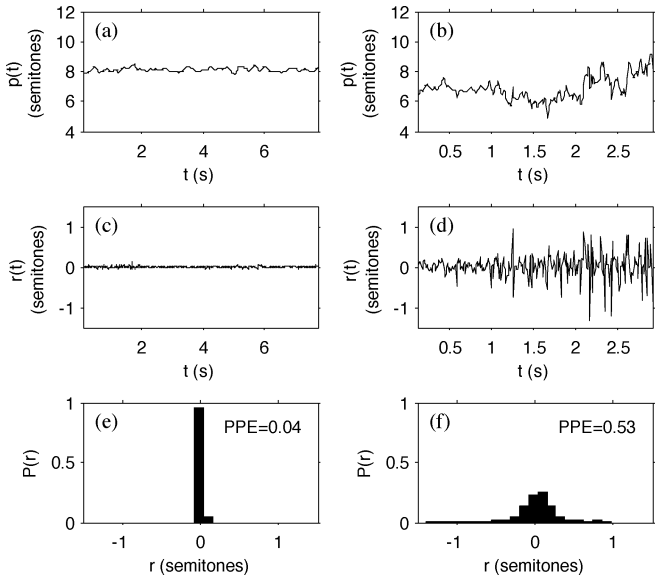

Fig. 3. Details of PPE calculation. (a) and (b) Pitch period p(t) in semitones relative to note C3 on the musical scale. (c) and (d) Residual of pitch period r(t) after spectral whitening filter. (e) and (f) Probability densities P(r) of residual pitch period r. PPE value is the entropy of this probability density). Left panels are for a healthy subject, right panel is for a person with Parkinson's.

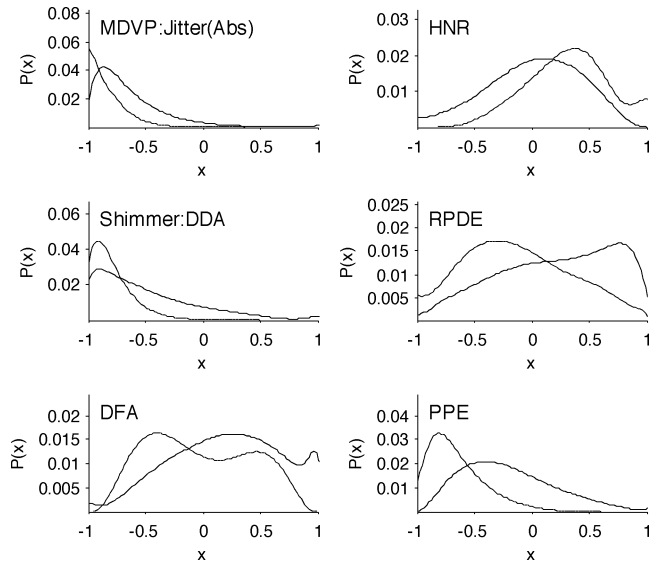

Fig. 4. Probability densities of some selected features after preprocessing by range normalization, in preparation for SVM classification (see Table II for a list of these features). The vertical axes are the probability densities P(x) of the normalized feature values x, estimated using the kernel density method with Gaussian kernel function. The dashed lines are for healthy subjects and the solid lines for Parkinson's subjects.

are very highly correlated and collinear, particularly the jitter and shimmer measures, whereas other measures are well spread relative to each other. This is particularly the case for the nonstandard measures or when comparing traditional with nonstandard measures. The correlation filtering removes the following features: MDVP:Jitter(%), MDVP:RAP, MDVP:PPQ, MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, and Shimmer:APQ5, leaving ten of the original measures (see Table II for a list of retained measures).

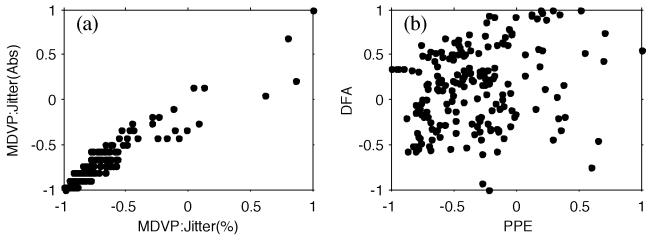

Fig. 5. Plots of pairs of features after preprocessing by range normalization, showing examples of high correlation (a) and low correlation (b). One of each pair of highly correlated features is removed prior to classification.

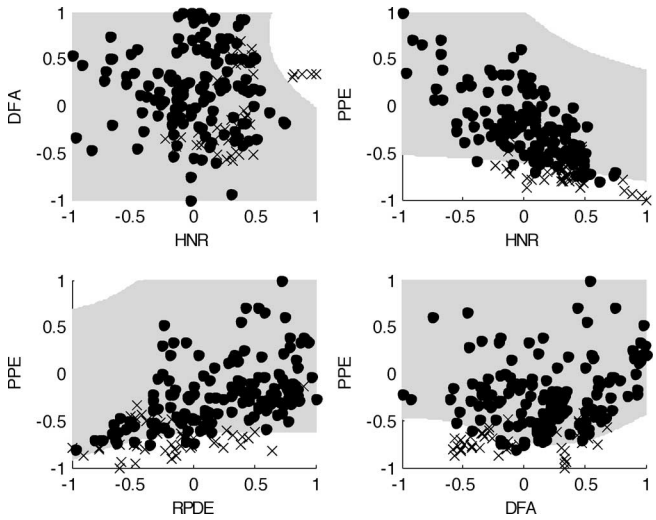

Fig. 6. SVM classification boundaries for some selected pairs of features after preprocessing by range normalization (see Table II for a list of these features). The "x'' marks are for healthy subjects and the round marks for Parkinson's subjects. The light gray shaded areas are the regions in which subjects are predicted to have Parkinson's.

The subsequent filtering of features leaves ten of the measures, and there are 1023 possible subsets of all these measures. It is therefore feasible to test all the combinations exhaustively. Table III details the resulting classification performance, with 95% confidence intervals, for some representative selected subsets of the measures retained after filtering. As can be seen, the combination of HNR, RPDE, DFA, and PPE obtains best overall classification performance, followed by the combination of all ten filtered measures. When taken separately, PPE produces the best performance. Fig. 6 shows the results of SVM classification applied to selected pairs of the four measures HNR, RPDE, DFA, and PPE. The boundaries are somewhat complex with some significant curvature. As can be seen, when PPE is included, the healthy and PD classes become better separated, and this is born out in the overall classification performance where the PPE measure contributes significantly toward a big improvement in the effectiveness of the classification.

V. Discussion

Our main finding is that nonstandard measures significantly outperform the traditional measures in separating healthy

controls from PWP, in terms of overall correct classification performance. We also find that traditional noise-to-harmonics methods contain some useful information that somewhat increases the performance. Furthermore, incorporating knowledge of and adjusting for the effect of natural pitch period variations leads to the design of a new measure PPE gaining significant performance increase. Considering the total number of signals is 195, 75.4% of the signals are from PWP; we can therefore consider this as a "null" rate. Any combination of measures that cannot achieve significantly better than this rate is not practically useful. When taken separately, of the traditional measures, only the retained jitter measure is able to achieve a rate much above this. By contrast, the PPE measure alone is comfortably above the null rate. We also find that the PPE measure appears in all the best performing subsets.

Another important observation is that simply increasing the combination subset size does not automatically lead to increasing overall classification performance. For the size of the data, the optimum number of measures is about four, above which or below which the classification performance is compromised. Of the nonstandard measures, we find that D2 is the least reliable. This is largely because many of the speech signals are noisy, and this spuriously increases the measured correlation dimension. This is an essential limitation of the usefulness of the algorithm for noisy signals [12], [45]. On this point also, it is well known that the traditional measures can only be applied to those cases where the signal is highly repetitive [46]. Nonstandard measures, other than D2, do not suffer from this limitation.

We believe that the results caution against the use of traditional measures of dysphonia for telemonitoring applications. The careful design and combination of novel, nonstandard measures, which are robust to variations in certain environmental conditions and to natural variations in individual voices, can lead to effective and reliable methods with which to discriminate healthy controls from PWP for remote monitoring applications. An important note is that our results are based on broadband, uncompressed audio signals, and we assume that future Internet bandwidth is sufficient that voice compression will not generally be required. Future research could further test these findings by applying these measures to voice signals recorded in acoustic environments more typical of practical telemonitoring applications.

ACKNOWLEDGMENT

The authors are grateful to M. Deisher and B. DeLeeuw at Intel Corporation and Athanasios Tsanas for comments on early drafts of the paper, and for the comments of the three anonymous reviewers that prompted improvements to the paper.

REFERENCES

[1] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, pp. 341–345, 2001.

[2] M. M. Hoehn and M. D. Yahr, "Parkinsonism—Onset progression and mortality," *Neurology*, vol. 17, 1967.

[3] A. E. Lang and A. M. Lozano, "Parkinson's disease—First of two parts," *New. Engl. J. Med.*, vol. 339, pp. 1044–1053, 1998.

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

[5] D. M. Huse, K. Schulman, L. Orsini, J. Castelli-Haley, S. Kennedy, and G. Lenhart, "Burden of illness in Parkinson's disease," *Movement Disord.*, vol. 20, pp. 1449–1454, 2005.

[6] N. Singh, V. Pillay, and Y. E. Choonara, "Advances in the treatment of Parkinson's disease," *Progr. Neurobiol.*, vol. 81, pp. 29–44, 2007.

[7] C. Ruggiero, R. Sacile, and M. Giacomini, "Home telecare," *J. Telemed. Telecare.*, vol. 5, pp. 11–17, 1999.

[8] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomed. Eng. Online.*, vol. 6, p. 23, 2007.

[9] 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, pp. 131–137, 1998.

[10] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and co-occurrence of vocal-tract dysfunctions in speech of a large sample of Parkinson patients," *J. Speech. Hear. Disord.*, vol. 43, pp. 47–57, 1978.

[11] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 2nd ed. St. Louis, MO: Elsevier, 2005.

[12] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, New ed. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[13] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of intensive voice treatment (the Lee Silverman voice treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings," *J. Speech. Lang. Hear. Res.*, vol. 50, pp. 899–912, 2007.

[14] D. A. Rahn, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory impairment in Parkinson's disease: Evidence from nonlinear dynamic analysis and perturbation analysis," *J. Voice*, vol. 21, pp. 64–71, 2007.

[15] K. M. Rosen, R. D. Kent, A. L. Delaney, and J. R. Duffy, "Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers," *J. Speech. Lang. Hear. Res.*, vol. 49, pp. 395–411, 2006.

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

[17] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. V. D. Heyning, M. Remacle, and V. Woisard, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the committee on phoniatrics of the European Laryngological Society (ELS)," *Eur. Arch. Otorhinolaryngol.*, vol. 258, pp. 77–82, 2001.

[18] J. Alonso, J. de Leon, I. Alonso, and M. Ferrer, "Automatic detection of pathologies in the voice by HOS based parameters," *EURASIP. J. Appl. Signal Process.*, vol. 4, pp. 275–284, 2001.

[19] M. Little, P. McSharry, I. Moroz, and S. Roberts, "Nonlinear, biophysically-informed speech pathology detection," in *Proc. ICASSP 2006*. New York: IEEE Publishers, 2006.

[20] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.

[21] S. Hadjitodorov, B. Boyanov, and B. Teston, "Laryngeal pathology detection by means of class-specific neural maps," *IEEE Trans. Inf. Technol. Biomed.*, vol. 4, no. 1, pp. 68–73, Mar. 2000.

[22] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices," *IEEE Eng. Med. Biol. Mag.*, vol. 16, no. 4, pp. 74–82, Jul./Aug. 1997.

[23] J. H. L. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 300–313, Mar. 1998.

[24] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Defebvre, and F. Grenez, "Low-frequency vocal modulations in vowels produced by Parkinsonian subjects," *Speech. Commun.*, vol. 50, pp. 288–300, 2008.

[25] P. Zwirner, T. Murry, and G. E. Woodson, "Phonatory function of neurologically impaired patients," *J. Commun. Disord.*, vol. 24, pp. 287–300, 1991.

[26] M. A. Little, "Biomechanically informed nonlinear speech signal processing," D.Phil. dissertation, Univ. Oxford, Oxford, U.K., 2007.

[27] J. J. Jiang and Y. Zhang, "Chaotic vibration induced by turbulent noise in a two-mass model of vocal folds," *J. Acoust. Soc. Amer.*, vol. 112, pp. 2127–2133, 2002.

[28] M. Little, P. McSharry, I. Moroz, and S. Roberts, "Testing the assumptions of linear prediction analysis in normal vowels," *J. Acoust. Soc. Amer.*, vol. 119, pp. 549–558, 2006.
[29] J. Zhang and M. Small, "Complex network from pseudoperiodic time series: Topology versus dynamics," *Phys. Rev. Lett.*, vol. 96, p. 238701, 2006.
[30] J. Zhang, X. Luo, and M. Small, "Detecting chaos in pseudoperiodic time series without embedding," *Phys. Rev. E.*, vol. 73, pp. 016216-1–016216-5, 2006.
[31] C. J. Huberty and L. L. Lowman, "Group overlap as a basis for effect size," *Edu. Psychol. Meas.*, vol. 60, pp. 543–563, 2000.
[32] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations*. New York: Springer-Verlag, 2001.
[33] P. E. McSharry, L. A. Smith, and L. Tarassenko, "Prediction of epileptic seizures: Are nonlinear methods relevant?," *Nat. Med.*, vol. 9, pp. 241–242, 2003.
[34] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
[35] J. Svec, P. Popolo, and I. Titze, "Measurement of vocal doses in speech: Experimental procedure and signal processing," *Logoped. Phoniatr. Vocol.*, vol. 28, pp. 181–192, 2003.
[36] KayPENTAX , "Kay elemetrics disordered voice database, model 4337," Kay Elemetrics, Lincoln Park, NJ, 1996–2005.
[37] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," presented at the Inst. Phonet. Sci., University of Amsterdam, Amsterdam, The Netherlands, 1993, vol. 17.
[38] J. J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement," *J. Voice*, vol. 20, pp. 2–17, 2006.
[39] R. Hegger, H. Kantz, and T. Schreiber, "Practical implementation of nonlinear time series methods: The TISEAN package," *Chaos*, vol. 9, pp. 413–435, 1999.
[40] R. P. Dixit, "On defining aspiration," in *Proc. 12th Int. Conf.*, Linguistics, Tokyo, Japan, 1988, pp. 606–610.
[41] J. Schoentgen and R. Deguchteneere, "Time-series analysis of jitter," *J. Phonet.*, vol. 23, pp. 189–201, 1995.
[42] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. Boston, MA: Academic, 2003.
[43] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 1996.
[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.
[45] Y. Zhang and J. J. Jiang, "Nonlinear dynamic analysis in signal typing of pathological human voices," *Electron. Lett.*, vol. 39, pp. 1021–1023, 2003.
[46] P. N. Carding, I. N. Steen, A. Webb, K. Mackenzie, I. J. Deary, and J. A. Wilson, "The reliability and sensitivity to change of acoustic measures of voice quality," *Clin. Otolaryngol.*, vol. 29, pp. 538–544, 2004.

Max A. Little (M'06) was born in Edinburgh, in 1971. He received the B.Sc. (Hons.) degree in mathematical sciences from the Open University, Milton Keynes, U.K., in 2003, and the D.Phil. degree in engineering science and applied mathematics from the University of Oxford, Oxford, U.K., in 2007.

He was a Researcher engaged in developing signal processing algorithms for several organizations including Creative Laboratories and a variety of video games companies. He is the holder of five patents for novel signal processing algorithms, one of which is used in Microsoft's XBox hardware. He is the Co-Founder of the Systems Analysis, Modeling and Prediction Group, University of Oxford. His current research interests include statistical signal processing, nonlinear time series analysis and machine learning techniques applied to complex physical systems, with applications in biological and environmental contexts.

Dr. Little is a member of the London Mathematical Society and the Performing Rights Society. He was the recipient of the Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award and a Prize Paper Award at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2006.

Patrick E. McSharry, photograph and biography not available at the time of publication.

Eric J. Hunter, photograph and biography not available at the time of publication.

Jennifer Spielman, photograph and biography not available at the time of publication.

Lorraine O. Ramig, photograph and biography not available at the time of publication.