Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease

Athanasios Tsanas, Max A. Little, Cynthia Fox, Lorraine O. Ramig

***Abstract*—Vocal performance degradation is a common symptom for the vast majority of Parkinson's disease (PD) subjects, who typically follow personalized one-to-one periodic rehabilitation meetings with speech experts over a long-term period. Recently, a novel computer program called Lee Silverman Voice Treatment (LSVT) Companion was developed to allow PD subjects to independently progress through a rehabilitative treatment session. This study is part of the assessment of the LSVT Companion, aiming to investigate the potential of using sustained vowel phonations towards objectively and automatically replicating the speech experts' assessments of PD subjects' voices as 'acceptable' (a clinician would allow persisting during in-person rehabilitation treatment) or 'unacceptable' (a clinician would not allow persisting during in-person rehabilitation treatment). We characterize each of the 156 sustained vowel /a/ phonations with 309 dysphonia measures, select a parsimonious subset using a robust feature selection algorithm, and automatically distinguish the two cohorts (acceptable versus unacceptable) with about 90% overall accuracy. Moreover, we illustrate the potential of the proposed methodology as a probabilistic decision support tool to speech experts to assess a phonation as 'acceptable' or 'unacceptable'. We envisage the findings of this study being a first step towards improving the effectiveness of an automated rehabilitative speech assessment tool.**

Manuscript received June 25, 2013; revised 19 October 2013; accepted 24 November 2013. A. Tsanas was funded by the Engineering and Physical Sciences Research Council (EPSRC), UK. His research is currently supported by the Wellcome Trust through a Centre Grant No. 098461/Z/12/Z, 'The University of Oxford Sleep and Circadian Neuroscience Institute (SCNi)'. M.A. Little acknowledges the financial support of the Wellcome Trust, grant number WT090651MF. This study was supported, in part, by NIH Grant No. 1R43DC010956-01. *Asterisk indicates corresponding author.*

*A. Tsanas is with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK, and with the Wolfson Centre for Mathematical Biology, Mathematical Institute, University of Oxford, Oxford, UK (phone: 0044 1865283874; fax: 0044 1865270515; e-mail: tsanas@maths.ox.ac.uk, tsanasthanasis@gmail.com).

M. A. Little is with the Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA (maxl@mit.edu).

C. Fox is with the Speech, Language, and Hearing Science, University of Colorado, Boulder, Colorado, USA, and with the National Center for Voice and Speech, Denver, Colorado, USA (cynthia.fox@lsvtglobal.com)

L. O. Ramig is with the Speech, Language, and Hearing Science, University of Colorado, Boulder, Colorado, USA and with the National Center for Voice and Speech, Denver, Colorado, USA (Lorraine.Ramig@colorado.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes an Excel file which is 34 Kb in size.

***Index Terms*—Decision support tool, Lee Silverman Voice Treatment (LSVT), Parkinson's disease, speech rehabilitation**

I. Introduction

PARKINSON'S DISEASE is a chronic neurodegenerative disorder characterized by the progressive deterioration of motor function and the emergence of considerable non-motor problems [1]. It is estimated there are at least 100 PD subjects per 100,000 in the population [2], and some studies have suggested PD prevalence may be underestimated [3]. Vocal impairment is reported in the vast majority of PD subjects, and approximately 29% of those consider it one of their greatest hindrances associated with the disease [4]. Moreover, speech performance degradation may be amongst the first symptoms of PD onset [5].

Typical vocal impairment symptoms include reduced loudness, monotone, hoarseness, breathiness (noise), imprecise articulation and vocal tremor [6]. The extent of vocal impairment can be assessed using sustained vowel phonations, or running speech. It can be argued that while some of the vocal deficiencies in running speech caused by PD (e.g. sequences of consonants and vowels) may not occur in sustained vowels, the analysis of running speech is inherently more complex due to articulatory and other linguistic confounds [7], [8]. Consequently, sustained vowel phonations, where the speaker attempts to produce a vowel sound as steady as possible (in terms of amplitude and frequency) and for as long as possible, are commonly used in clinical practice [8]. Both clinical practice and extensive research have shown that the sustained vowel "ahh…" (denoted /a/) may be sufficient for many voice assessment applications [8], [9], [10], [11], and for PD voice assessment in particular [12], [13], [14], [15], [16], [17].

Clinical speech signal processing algorithms (algorithmic tools extracting clinically useful information from speech signals) are collectively known as *dysphonia measures*. From the point of view of speech experts and ear, nose and throat surgeons, there is an important distinction between *voice* (the sound produced by the larynx) and *speech*, which is the integrated process of voice and articulation. Despite the subtle difference in the narrow definition of voice and speech, I. Titze asserts "*in the broader sense voice is synonymous with speech*" [8]. Strictly speaking, the dysphonia measures discussed in this study are entirely based on voice rather than speech since we do not attempt to characterize articulation problems.

LSVT LOUD is a standardized, research-based speech treatment protocol with established efficacy for PD [18], [19],

[20]. A significant challenge is how to scale accessibility to this speech treatment program that requires a sustained, intensive treatment regime when there are not enough clinicians to deliver all the treatment that is needed. Advances in computer and web-based technology offer solutions to the problems of treatment accessibility, efficacious treatment delivery, and long-term maintenance in rehabilitation [21]. Such technology may alleviate the barrier of inadequate numbers of clinicians to deliver in-person therapy, enhance the feasibility of delivering intensive treatment requirements, and relieve the logistical burden of traveling to and from the clinic for in-person treatment. Through funding from the National Institutes of Health (NIH) and the Michael J. Fox Foundation, LSVT Global has developed and tested the *LSVT Companion*, a computer program that allows PD subjects to independently progress through a treatment session. The LSVT Companion automatically obtains data on sound pressure level (SPL), fundamental frequency (F0), and phonation duration during voice tasks, and provides instantaneous audio and visual feedback to the client on their vocal loudness, pitch and phonation duration during treatment exercises. When PD subjects do not meet their target goals (vocal loudness, pitch, duration), the LSVT Companion provides feedback to encourage them to appropriately adjust the characteristics of their phonation, e.g. louder, longer, higher or lower.

In a recent study, the LSVT Companion was used to augment in-person treatment sessions [22]. Seven of the 16 sessions were completed with subjects using the LSVT Companion for independent at-home therapy and nine of the sessions were traditional in-person sessions. Outcome data immediately post-LSVT and six months later for SPL were comparable to data from 16 in-person treatment sessions with a clinician. The next step in LSVT Companion development is to further support independent treatment delivery (i.e., increase the number of independent at-home sessions). One potential concern is that in the process of learning to improve vocal loudness, PD subjects may exhibit unacceptable voice characteristics that an LSVT expert clinician would "not allow to persist" during in-person treatment.

The aim of the current study is to investigate the potential of using an objective statistical machine learning framework to automatically evaluate sustained vowel phonations as "acceptable" (a clinician would allow persisting in speech treatment) or "unacceptable" (a clinician would not allow persisting in speech treatment). The ultimate goal is to improve the effectiveness of rehabilitative speech treatment by developing an appropriate algorithm for the LSVT Companion system. This algorithm will be able to detect unacceptable voice characteristics during use of software away from expert clinical guidance, stop the patient from using voice in an unacceptable way, and subsequently improve voice characteristics through providing feedback.

II. DATA

Seventeen subjects with PD diagnosis were originally screened for inclusion in this study, of whom 14 finally participated. The three subjects who did not participate in this study were due to scheduling conflicts with the data collection sessions. All subjects had typical voice and speech characteristics of PD as determined by an experienced speech-language pathologist (e.g. reduced loudness, monotone, breathy, hoarse voice, or imprecise articulation) upon telephone screening, and verified by two speech-language pathologists with expertise in PD during data collection. The subjects' enrolment in this study and all recruiting materials were approved by an independent Institutional Review Board.

The 14 PD subjects (8 males and 6 females), had an age range of 51 to 69 (mean ± standard deviation: 61.9 ± 6.5) years, and produced sustained vowel /a/ phonations. The sustained vowel phonations were recorded in a double-walled, sound-attenuated room at the National Center for Voice and Speech-Denver (NCVS), an affiliate institution of the University of Colorado-Boulder. A head-mounted microphone was positioned according to standard protocols [23]. The voice signals were sampled at 44.1 kHz with 16 bits of resolution, and were recorded using the Audacity software package (http://audacity.sourceforge.net/). In total, each subject was originally instructed to produce 27 phonations (samples) where each phonation belonged to one of the nine possible combinations of pitch and amplitude, that is, phonations at comfortable pitch, high pitch, low pitch, with amplitude considered acceptable, too loud or too soft. The rationale behind using different conditions of pitch and amplitude is to have samples from diverse types of phonation; research has shown that consciously thinking of the way we speak may result in vocal performance improvement [24]. The aim was to use one good sample from each of the nine combinations: in practice it took more than a single attempt to record good samples (some subjects laughed, or coughed). Similarly, it took more than one attempt for some subjects to follow the instruction regarding the task they were required to complete (for example when stimulating high amplitude phonations). Then, for each subject, the data collection team selected (prior to any speech signal analysis) the best sample for each of the nine possible combinations of pitch and amplitude, and assessed each sample as "acceptable" or "unacceptable". This data selection step was to ensure that we would process the same number of samples for each subject, and using one good representative sample for each combination of pitch and amplitude for each subject avoids contaminating the dataset with phonations which may not be sufficiently "good". The discarded phonations did not meet target criteria (e.g. not loud enough for the task "high amplitude", no change in pitch when instructed to produce low or high pitch phonations, cough or laugh, etc.). Thus, we processed nine phonations per subject for a total of 126 phonations for the 14 PD subjects. Twenty-five percent of those 126 samples were repeated to quantify *intra-rater reliability* for a total of 156 samples.

LSVT expert clinicians assessed *perceptually* whether phonations could be considered "acceptable" or "unacceptable". There are no specific objective criteria with which speech experts assess phonations; the assessment largely depends on the experience of the rater. Each subject produced phonations of both types ("acceptable" and "unacceptable"), as assessed by consensus by two experienced speech experts (C.F. and L.R.). This assessment will be considered the 'ground truth'. We also have five additional assessments per phonation from five different speech experts, who were blinded to all the other raters' assessments and the

subjects' condition. In all cases the expert raters had to decide whether each phonation was "acceptable" or "unacceptable". Therefore, we have six assessments from expert raters for each of the 156 samples.

Each phonation was pre-processed, selecting a 2-second segment from the stable part of the phonation (middle of the speech signal) for subsequent acoustic analysis in order to avoid problems during the onset (start) and offset (end) of the phonation. Manual inspection did not reveal any problematic recordings (anything but an 'ahh…' sound, e.g. coughing), so all data samples were used in the analysis.

III. Methods

The aim of this study is to process the speech signals, extract information-rich dysphonia measures (generally referred to as *features* in machine learning), and map the most *parsimonious* feature subset (feature subset which is maximally informative with as few features as possible) to the *response* (acceptable versus unacceptable).

A. Feature calculation

We apply the dysphonia measures defined in detail in Tsanas *et al*. [14], [25], and summarized more recently in Tsanas [26]. Many dysphonia measures rely on the computation of the fundamental frequency (F0). In this study, we use the Non-Defect-Free (NDF) F0 estimation algorithm [27]: it was recently shown to consistently outperform popular competing F0 estimators in the context of sustained vowel /a/ phonations [26], on comparison against reference, synthetic speech data created by a state of the art physiological model. Here, we summarize the dysphonia measures used in this study, clustering them into groups. We refer to Tsanas [26] and references therein for details and the rationale behind each dysphonia measure.

The first group of dysphonia measures builds on the physiological observation that the vocal fold vibration pattern is nearly periodic in healthy voices (type I according to Titze [8]), whereas pathological voices tend to depart from periodicity or are completely aperiodic (types II and III according to Titze [8]). Two of the most widely used dysphonia measures fall under this category, and are known as *jitter* and *shimmer* [7], [8]. Both jitter and shimmer are classical perturbation measures: jitter quantifies F0 deviations, whereas shimmer quantifies deviations in amplitude. There is no unique mathematical definition of jitter and shimmer, therefore we investigated many *jitter variants* and *shimmer variants*, which are algorithmic variations of the same basic ideas. The concept of quantifying deviations in the vocal fold vibration pattern has inspired us to propose the *Recurrence Period Density Entropy* (RPDE) [28], the *Pitch Period Entropy* (PPE) [12], the *Glottal Quotient* (GQ) [14], and other F0-related measures [14]. RPDE quantifies the uncertainty in estimating the duration of the vocal fold cycle. PPE quantifies impairments in controlling F0 in sustained vowel phonations. GQ is very similar to jitter measures, but operates on vocal fold cycles instead of adjacent time segments of the speech signal. The F0-related dysphonia measures include statistical summaries of the F0 distribution, and the difference in the measured F0 of age- and gender-matched healthy controls.

The second general group of dysphonia measures is the Signal to Noise Ratio (SNR) type algorithms. The physiological motivation for this group is that incomplete vocal fold closure leads to the creation of aerodynamic vortices which result in increased acoustic noise. *Harmonic to Noise Ratio* (HNR) [7], *Detrended Fluctuation Analysis* (DFA) [28], *Glottal to Noise Excitation* (GNE) [29], *Vocal Fold Excitation Ratio* (VFER) [14], and *Empirical Mode Decomposition Excitation Ratio* (EMD-ER) [14] are archetypal examples of this group. GNE and VFER analyze frequency ranges of sustained vowel phonation in bands of 500 Hz. We have empirically found in previous studies that frequencies below 2.5 kHz can be treated as 'signal', and everything above 2.5 kHz can be treated as 'noise' [14], [26]. Thus, we can define SNR dysphonia measures using energy, nonlinear energy (Teager-Kaiser energy operator) and entropy concepts. Organic pathology-free voices may be harmonically efficient up to 6-7 kHz (although this might not be true for the elderly), and hence the use of 2.5 kHz as the threshold for 'noise' might need further clarification. We first reported on this empirical finding for considering frequencies below 2.5 kHz as 'signal' and frequencies above 2.5 kHz as 'noise' in Tsanas *et al*. [14], where the threshold was optimized in order to determine a PD symptom severity clinical scale. The 2.5 kHz threshold has a tentative physiological justification: most of the energy in the sustained vowels is contained mainly up to the second formant, and for the sustained vowel /a/ the second formant can be up to about 1.7 kHz [8]. Interestingly, another research group has reported that frequencies above 2 kHz can be generally considered turbulent noise [30]. Similarly, the Multi-Dimensional Voice Program (MDVP - http://www.kayelemetrics.com/) uses a dysphonia measure called "Voice Turbulence Index", where the spectral frequencies above 2.8 kHz are used to quantify the high frequency energy component in the speech signal [31]. EMD-ER has a similar conceptual basis forming ratios of signal and noise energies. The empirical mode decomposition (EMD) algorithm [32] decomposes a signal into elementary, linearly superposed signal components with amplitude and frequency contributions. Then, the top (high frequency) components are taken to constitute noise, whereas the lower frequency components are taken to constitute the signal. We defined the EMD-ER dysphonia measures using similar ideas as in VFER, i.e. energy and entropy concepts.

The wavelet measures proposed in Tsanas *et al*. [25] form a generic approach to analyzing time series signals: we suggested using wavelet decomposition with ten levels of decomposition, to analyze the F0 time series (also known as F0 *contour*). The wavelet coefficients are then the features presented to the classifier (see section III.D). Inspired by previous work [12], [33] we also used the log-transformed F0 time series, because this power transformation can sometimes help bring out additional characteristics in dysphonia measures [33].

Lastly, *Mel Frequency Cepstral Coefficients* (MFCCs) have been widely used in automatic speech recognition and speaker identification, and recent studies have shown that they are promising in biomedical applications as well [14], [15], [26], [34]. MFCCs target the placement of the articulators (collectively referring to the mouth, teeth, tongue, and lips),

which is known to be affected in PD [35]. We clarify that the articulatory features used in this study are characterizing fluctuations and instability in postural stability of articulators during sustained vowel phonation, and are not used to characterize dysarthria.

Overall, we calculated 309 dysphonia measures using the speech database, where each dysphonia measure produced a single number per phonation, resulting in a design matrix of size 156×309. There were no missing entries in the design matrix.

B. Data exploration and statistical analysis

Exploratory analysis is typically the first step in most applications involving dealing with data, in order to gain a preliminary understanding of the statistical properties of the dataset. It includes data visualization plots (e.g. scatter plots) and the computation of statistical associations [36]: here, we computed the Pearson correlation coefficient and the normalized mutual information [14] to assess the association strength between each feature and the response. The normalized mutual information ranges from 0 to 1, where larger values indicate stronger statistical association between the feature and the response.

C. Feature selection

The very large number of dysphonia measures used in this study (309) may lead to overfitting due to the curse of dimensionality: the feature space will be inadequately populated with only 126 samples. Recent findings suggest that even powerful classifiers such as support vector machines and random forests are not immune to the curse of dimensionality, and their performance may degrade as a result of including too many features [15]. Reducing the number of features may or may not improve the generalization performance of the classifier (i.e. performance in a novel dataset); however a reduced feature subset typically facilitates insight into the problem via analysis of the most predictive features [37]. Exhaustive search through all possible feature subsets is computationally infeasible. Instead, *feature selection* (FS) algorithms offer a principled, computationally tractable approach to select a parsimonious, information-rich feature subset. Contrary to feature transformation algorithms (the most popular example is principal component analysis, for details see Bishop [38]), which transform the original features to build a new feature set with new properties, in FS the results are more interpretable because we retain domain expertise. Hence, we try to determine "which of the originally computed dysphonia measures really matter in this problem".

Here, we used a recently proposed FS algorithm, LOGO (fit locally and think globally), which has shown promising results over a wide range of different applications [39]. LOGO is a feature weighting algorithm, where each feature is assigned a weight which can be interpreted as the 'importance' of the feature in predicting the response in the presence of the other features. The feature subset was selected using cross-validation (CV), using only the training data at each CV iteration (90% of the data, randomly selecting samples at each iteration, see section III.E for details). The CV process was repeated a total of 100 times, where in each iteration the M features (M=309) appear in descending order of selection. Ideally, the FS and ranking should be identical for all CV iterations, however in practice this is often not the case. In order to identify the feature subset we use the strategy previously outlined in Tsanas [15], [26]. Specifically, we create an empty set S which will contain the indices of the selected features. We apply the following voting scheme, where feature indices are incrementally included, one at a time, in S. For each step K (K=1…M) we find the indices corresponding to the features selected in the 1…K search steps for the 100 CV iterations. We select the feature index which appears most frequently amongst the 100×K elements that has not been previously included in S. This index is now included as the Kth element in S. Ties are resolved by including the lowest index number. Moreover, we introduce the *stability weight*, which is the percentage of times in the 100 CV iterations that each feature was selected in the feature subset containing K features: this expresses the confidence of the FS algorithm in selecting each feature in perturbed versions of the dataset (which are obtained by randomly selecting 90% of the samples in each iteration and repeating the process 100 times). The larger the stability weight, the more robust the results of the FS algorithm, and the more confident we are the feature is indeed predictive of the response.

The ranked feature subsets can now be presented into the classifier in the subsequent mapping phase to estimate the binary response "acceptable" or "unacceptable" using the dysphonia measures.

D. Statistical mapping: estimating the response

Although correlation analysis may give an indication of the association strength of each feature with the response and the potential of differentiating the two classes in this study, we need to build a functional relationship f(X)=y, which maps the dysphonia measures X to the response y. That is, we need a *binary classifier* that will use the dysphonia measures to discriminate phonations as "acceptable" or "unacceptable". We compared two widely-used statistical machine learning algorithms here: *Random Forests* (RF) [40], and *Support Vector Machines* (SVM) [41].

RF is an ensemble method, weighting the output of a large number of base learners (random decision trees). Typically, the RF output regarding the class of a new sample is simply the class that gets the majority of votes from the base learners. RF is fairly insensitive to the choice of the free parameters [40]: (1) the number of trees, which should be fairly large (we used 500 trees, which is the default suggestion), and (2) the number of features over which to search to construct each branch of each tree (we used the square root of the input feature space, which is again the default suggestion).

In many practical applications, it may be useful to predict the response in a probabilistic setting. That is, a new unseen data sample is assigned a probability belonging to each of the possible classes, rather than a single class. The larger this probability is for one of the classes, the more confident we are that the sample in fact belongs to that class. With RF we can use the proportion of the output of the decision trees to determine the probability that a new sample belongs to each of the possible classes.

SVM rely on the margin maximization principle, constructing an optimal separating hyper-plane in the feature

space. The aim is to maximize a geometric margin between points from the two classes. In most practical applications data will not be linearly separable; in those cases SVM transform the data into a higher dimensional space, and construct the separating hyperplane in the new space [38]. SVM are known to be particularly sensitive to the values of their free parameters, and great care needs to be exercised when optimizing this classifier. Here, we used a Gaussian radial basis function kernel and linearly scaled each feature to lie in the range [-1, 1], which is the standard approach [42]. 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 III.E for details). Specifically, we searched over the grid (C, γ) defined by the product of the sets: $C = [2^{-5}, 2^{-13},\ldots, 2^{15}]$, and $\gamma = [2^{-5}, 2^{-13},\ldots, 2^{3}]$. Once the optimal parameter pair (C, γ) was determined, we trained and tested the classifier using these parameters. We used the LIBSVM implementation [42].

Similarly to RF, it is possible to obtain probabilistic outputs with SVM. We refer to Wu *et al*. [43] for an overview of the different strategies to obtain probability estimates with SVM. Here, we used the default LIBSVM strategy, which relies on a sigmoid function to determine the probability that the query sample belongs to either of the two possible classes in the binary SVM setting [43].

E. Classifier validation and generalization performance

The generalization performance of the classifier is an estimate of the performance we might expect on a novel feature dataset, assuming this novel dataset will come from a similar distribution to the feature data used to train the classifier. Typically, this validation is achieved using CV or bootstrap techniques [44]. Here, we used a 10-fold CV scheme, where the original data samples (126 phonations) were split into two subsets: a training subset consisting of 90% of the data samples (113 phonations), and a testing subset consisting of 10% of the data samples (13 phonations). The process was repeated 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 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 and the classifier accuracy was computed as 1-MAE.

In addition, we report our findings when training the classifier using samples from 13 PD subjects and testing the performance of the classifier on the samples of the 14th subject (i.e. leave all the samples from one subject out of the training set and use them for testing – LOO testing scheme). The process is repeated for all 14 subjects, each time leaving the phonations of one subject out of the training set, and using those phonations to calculate the classifier accuracy.

F. Contaminating speech signals with additive white Gaussian noise or pink noise

The speech signals in this study were collected under controlled acoustic conditions, which might not resemble accurately the acoustic conditions where the LSVT Companion is used at home. Therefore, in addition to analysing the high quality speech signals, we also simulated more realistic ambient environments to obtain speech signals contaminated with noise. The aim is to investigate how well the proposed methodology might generalize in settings where the speech signals are recorded in a quiet room at the subjects' homes. Specifically, each of the 126 speech signals was contaminated with (a) additive white Gaussian noise, or (b) pink noise [45]. Both additive white Gaussian noise and pink noise are often used in diverse engineering applications to simulate noise. For example, pink noise has been observed in many electronic devices [45], and is often used to simulate realistic noisy environments over which vocal performance is assessed [46]. In both cases we experimented with relatively low power noise, because the head-mounted microphone in the LSVT Companion and the requirement of the LSVT treatment sessions to be completed in a quiet room should practically suggest that the contaminating noise has relatively low power compared to the actual speech signal (we used SNR=60 dB and SNR=40 dB).

IV. RESULTS

This section presents the main findings: firstly, based on the assessment of the phonations from the two most experienced raters (C.F and L.R.), after this, we investigate the automated probabilistic assessment of the speech samples to see how well they match the assessments of all raters. In section IV.A we use the 126 samples, and in the following sections we also introduce the additional (repeated) 30 samples to study intra-rater and inter-rater variability.

A. Analysis using the assessment of the most experienced raters

Table I presents the 10 dysphonia measures most strongly associated with the response, which was defined to be 0 for "acceptable" phonations and 1 for "unacceptable" phonations. All these dysphonia measures were statistically significantly correlated (p<0.01) with the response. The interpretation of the association strength between two random variables (here between each feature and the response) based on the magnitude of the correlation coefficient depends on the application and there are no strict guidelines; nevertheless, correlation coefficients with an absolute value larger than 0.3 can be considered relatively strong in medical applications [36], [47], [48]. These statistical findings suggest that the binary classification task in this study may be successful (i.e. we may expect to get relatively high accuracy). In addition, Fig. 1 presents density estimation plots of the 10 dysphonia measures most strongly associated with the response, providing a visual impression of the distribution of their values for the two classes.

TABLE I
STATISTICAL ASSOCIATIONS OF THE DYSPHONIA MEASURES WITH THE RESPONSE

Dysphonia measure	Description	Correlation coefficient	Mutual information
$IMF_{NSR,SEO}$	Noise to signal energy ratio of speech signal constituents (components)	-0.508	**0.474**
3rd MFCC	3rd MFCC coefficient	-0.368	**0.432**
2nd MFCC	2nd MFCC coefficient	-0.530	**0.404**
1st MFCC	1st MFCC coefficient	-0.502	**0.393**
DFA	Characterizes the extent of turbulent noise	-0.468	**0.386**
0th MFCC	0th MFCC coefficient	0.485	**0.379**
$IMF_{NSR,entropy}$	Noise to signal entropy ratio of speech signal constituents (components)	-0.487	**0.353**
$Jitter_{F0abs,diff}$	Mean absolute difference of successive fundamental frequency estimates	0.312	**0.319**
Log energy	Logarithm of the energy of the speech signal	0.419	**0.304**
$VFER_{NSR,TKEO}$	Extent of noise in speech using the nonlinear energy operator	-0.281	**0.303**

The ranking was determined by the mutual information (MI), and for clarity we present only the 10 dysphonia measures with the largest MI value. The reported MI is normalized (that is, it lies in the range 0 to 1, where 0 denotes that the response is independent of the dysphonia measure, and 1 indicates that the response is completely determined by the dysphonia measure). All results were rounded to the nearest third decimal digit. All dysphonia measures were statistically significantly correlated ($p < 0.01$) with the response. All samples were used to generate these results ($N = 126$ phonations). The response refers to the assessment of the phonation as "acceptable" or "unacceptable".

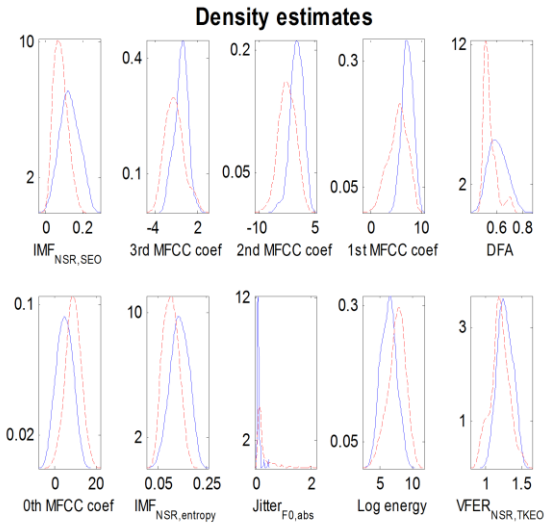

Fig. 1. Density estimation plots of the 10 most strongly associated dysphonia measures with the response (see Table 1). The red dotted lines correspond to pathological voices. The estimates were computed using kernel density estimation with Gaussian kernels.

Next, we use the LOGO FS algorithm to determine a parsimonious feature subset, which is presented in Table II along with the LOGO weight and the stability weight. It is interesting that many of the non-standard dysphonia measures (i.e. dysphonia measures other than jitter and shimmer, here mainly energy-related dysphonia measures) appear to be consistently selected by LOGO. The top five features in particular, appear to exhibit relatively large feature weights and are constantly selected (stable) in all 100 iterations of perturbed versions of the original dataset, which suggests these features may generalize well in new unseen data with the same properties. We note that a dysphonia measure which has larger weight or is selected before other dysphonia measures by LOGO does not necessarily mean it is more stable. Each of the LOGO weights across the 100 iterations (particularly the weights for the top six features) exhibited a quasi-Gaussian distribution (results not shown). LOGO weights where the mean value is larger than the standard deviation can be considered unstable (i.e. their weight value cannot be trusted). A very low weight value reflects that the feature does not

TABLE II
INCREMENTAL SELECTION OF A ROBUST FEATURE SUBSET USING THE LOGO ALGORITHM

Dysphonia measure	LOGO Weight	Stability weight (%)
Entropy detail wavelet coef 4th level of F0	5.05 ± 0.86	100
$IMF_{NSR,SEO}$	3.16 ± 0.92	100
0th MFCC	2.84 ± 0.75	100
DFA	2.26 ± 0.94	100
$VFER_{SNR,SEO}$	1.85 ± 0.64	100
$Shimmer_{A0,p1}$	0.40 ± 0.75	80
1st MFCC	0.18 ± 0.37	88
HNR_{std}	0.03 ± 0.22	38
$VFER_{NSR,TKEO}$	0.00 ± 0.00	58
$IMF_{NSR,entropy}$	0.00 ± 0.02	68

For clarity we only present the top 10 features selected by LOGO in descending order of importance. The LOGO weight is the weight of the LOGO feature selection algorithm: larger indicates the feature contributes more towards predicting the response. The LOGO weight is summarized in the form mean±standard deviation using the LOGO weights from 100 iterations, where each iteration uses a randomly selected 90% of the data (i.e. perturbed versions of the dataset). Each of the LOGO weights across the 100 iterations (particularly the weights for the top six features) exhibited a quasi-Gaussian distribution. The stability weight expresses how often each feature was selected in a 10-fold cross validation setting with 100 iterations: larger indicates that the feature was selected in more iterations, hence giving confidence this feature is predictive of the response. The weights have been rounded to the closest second decimal digit.

contribute towards predicting the response (for example the LOGO weight of $VFER_{NSR,TKEO}$ appears as 0.00±0.00 because we rounded all values to the closest second decimal digit and practically contributes no predictive information).

We compute the out of sample performance of SVM and RF as a function of presenting to the classifiers the top K features (K=1…30) selected using the LOGO FS algorithm. Figure 2 presents the performance results when using 10-fold CV with 100 iterations for statistical confidence, and Fig. 3 the results when training the SVM with the samples from 13 subjects and testing on the samples from the 14th subject (LOO testing). In both cases SVM was slightly better than RF, and hence the performance of RF is not shown. We observed that when at least eight features are used, the performance of the classifier fluctuates around 90% depending on the number of features presented to the classifier. The number of selected features can be decided on the basis of the results in Table II,

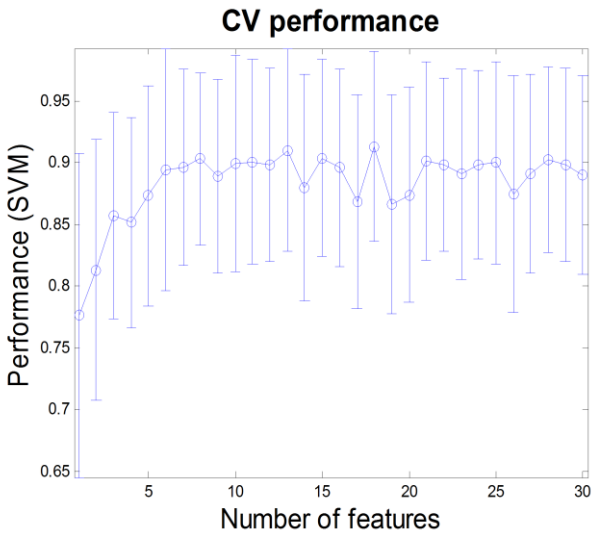

Fig. 2. Comparison of out of sample mean performance results with confidence intervals (one standard deviation around the quoted mean performance) as a function of the number of features, using the features selected using LOGO. These results are computed using 10-fold cross validation (CV) with 100 repetitions. For clarity, we present here only the first 30 steps of the feature selection algorithms.

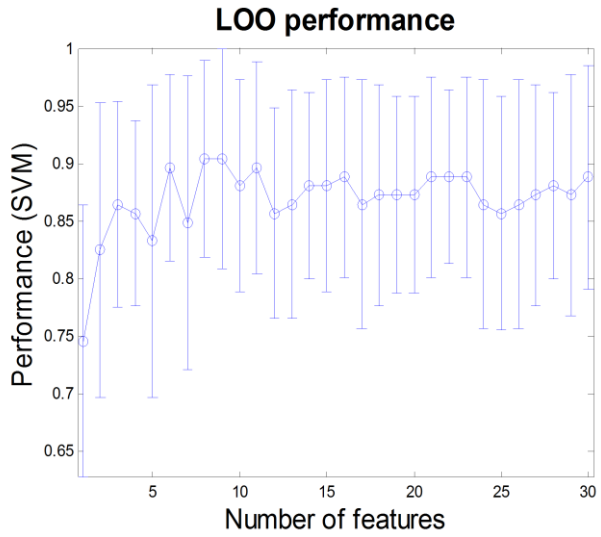

Fig. 3. Comparison of out of sample mean performance results with confidence intervals (one standard deviation around the quoted mean performance) as a function of the number of features, using the features selected using LOGO. These results are computed by training the classifier using the samples from all but one subject (leave one out – LOO), and testing the classifier's performance on the samples from the subject which were left out in the training process. This process is repeated for all 14 subjects and the results are averaged. For clarity, we present here only the first 30 feature selection steps.

and also verified by the findings in Fig. 2 and Fig. 3. We choose to use the top eight dysphonia measures in Table II: this satisfies the goal of accurate performance, and the rather more subjective goal of parsimony (selecting a small feature subset with good performance).

The results presented so far used the high quality speech signals collected under controlled acoustic conditions. To assess the robustness of the proposed methodology under more realistic noisy ambient environments, each of the 126 speech signals was contaminated with (a) additive white Gaussian noise, or (b) pink noise (see III.F). Subsequently, we repeated the methodology described previously for the characterization and mapping of the noisy signals to the response. In both cases we found that the classification accuracy was still approximately 90% when using the top eight dysphonia measures presented in Table 2.

B. Intra-rater variability

The dataset in this study consists of 126 samples and 30 additional repeated samples which were introduced to evaluate *intra-rater variability*. The inclusion of those 30 repeated samples was known in advance only to the most experienced expert rating team (C.F. and L.R.); hence, here we wanted to determine how well the five additional raters performed. The intra-rater reliability across the five expert raters was 100%, 97%, 94%, 90% and 84%. This was computed by averaging the proportion of times each of the five raters agreed with their own assessments in the 30 repeated samples.

C. Inter-rater variability: analysis using the assessment of all six raters

Of particular interest are those phonations where there is no complete agreement amongst experts (*inter-rater variability*), that is, when not all expert raters agree on the assessment of the phonation as "acceptable" or "unacceptable". It is not clear how to resolve this issue since there is no 'ground truth'; thus, the assessment obtained by consensus from the two most experienced raters (C.F. and L.R.) was used as the ground truth. The inter-rater reliability where the five expert raters agreed with the assessments of C.F. and L.R. across all samples was 50%, i.e. there was complete agreement in 78 out of the 156 phonations. The inter-rater reliability when using "majority rule" voting (three or more raters agreed with the assessment by C.F. and L.R.) was 85% across all phonations. In all cases where there was not complete agreement, one or more of the expert raters indicated the phonation was "difficult" to categorize.

Next, we wanted to investigate whether the statistical machine learning framework used in the study could provide some *probabilistic* guidance in the assessment of the 78 samples which are "difficult" to categorize. This could be a useful guide in practice, suggesting to clinical experts how confidently the algorithm assigns the phonation to either class. We used the 78 samples where there was complete agreement amongst all expert raters to train the classifiers, and interrogated the classifiers to provide *probabilistic* outputs. That is, both RF and SVM provided the probability that each phonation can be considered "acceptable" or "unacceptable". The results are summarized in the Excel file 'LSVTexpert_disagreement.xls' in the Electronic Supplementary Material. In some cases the probabilistic assessment is heavily driven towards one of the two classes, but there are a few samples where the algorithm assessed the output only marginally in favour of one of the two classes (in some cases the classifiers assigned the phonation to one of the two classes with less than 60% probability). Interestingly, most misclassifications (hard-thresholding the probabilistic outputs) of SVM and RF are precisely in those borderline

cases. We remark that those phonations appear to be particularly difficult to be correctly classified.

V. Discussion

Analysis of speech signals using clinical speech signal processing has generated considerable research interest in the last decade. This is, in large part, motivated by the great potential of speech as an information-rich signal: many studies have demonstrated we can extract clinically useful information in a variety of applications [26], [34], [49], [50], [51]. Moreover, speech signals are easy to collect since all they require is a microphone and a digital recording device that are, these days, ubiquitous in everyday life. In this study, we have applied an extensive pool of speech signal processing algorithms in order to investigate how accurately we can discriminate sustained vowel /a/ phonations which a clinician would allow to persist in speech treatment ("acceptable"), from those that a clinician would not allow to persist in therapy ("unacceptable"). The objective automatic characterization of acceptable and unacceptable voices is an essential step toward the ultimate goal of advancing the use of treatment software by PD subjects, independent of the clinic and expert staff. We demonstrated that we can achieve about 90% accuracy using a feature subset consisting of eight dysphonia measures. We remark that we did not observe substantial difference in the classifier accuracy when evaluating the proposed methodology using standard ten-fold CV with 100 iterations, or iteratively training the classifier with the samples from 13 subjects and testing its performance on the samples of the 14th subject (i.e. leave one subject out) which were not used in the training of the classifier (see Fig. 2 and 3).

We used a recently proposed robust FS algorithm to determine the most parsimonious feature subset: that is, the minimum set of dysphonia measures that jointly carries the maximum information towards predicting the response. The features were selected using a voting mechanism from 100 perturbed versions of the original dataset (each version using a randomly selected 90% of the 126 samples). We quantified confidence in the selected feature subset using the concept of feature stability: the selected features (particularly the top five features), are consistently selected on perturbed versions of the dataset. This finding strongly suggests the selected feature subset may generalize well in new unseen data with the same joint distribution as the dataset studied here.

As with previous studies we have found that the combination of classical ($Shimmer_{A0,p1}$), newly-proposed nonlinear speech signals processing algorithms ($IMF_{NSR,SEO}$, $VFER_{SNR,SEO}$), and low-order MFCCs may be the most information-rich feature subset that leads to accurate replication of the speech experts' assessment of the phonation (see Table II). The dysphonia measure *Entropy detail wavelet coef 4th level of F0*, which quantifies the entropy of the 4th decomposition level detail wavelet coefficient, is consistently selected as the most predictive feature with a large LOGO weight, indicating that wavelet decomposition of the fundamental frequency time series [25] is worthy of more detailed investigation in other speech signal processing applications, particularly when high-quality F0 estimates are available [26]. Also, this study reinforces the findings of previous studies that MFCCs appear to carry clinically useful information in the assessment of vocal disorders [14], [15], [34]. It is not easy to physically interpret the MFCCs, but the low-order MFCCs broadly express amplitude and formant fluctuations. Also, $IMF_{NSR,SEO}$, DFA, and $VFER_{SNR,SEO}$ are SNR-type dysphonia measures quantifying excessive noise in the phonation. Collectively, these dysphonia measures suggest that incomplete vocal fold closure and vocal tremor which lead to the creation of aerodynamic vortices resulting in increased acoustic noise, may be the primary characteristic that clinicians use to perceptually assess whether a phonation is considered "acceptable" or "unacceptable".

Recent studies have suggested that it may be useful to partition the data according to gender [14], [26], [52]. Potentially, this can offer insight into differences of voices observed in males and females and possibly indicate something about gender-dependent characteristics. However, splitting the original dataset into two subsets reduces the statistical power of the performance evaluations, because of the relatively limited number of samples available to this study. This was verified when we partitioned the data according to gender, and obtained reduced performance accuracy. It is possible that using a larger dataset and partitioning the data might lead to interesting new insights, and increased overall performance accuracy. For example, data partitioning by gender provided tentative physiological insight into the differences of the pathophysiological mechanism in PD for males and females [14], [26].

The voice samples of this pilot study were collected under high quality controlled conditions: this was the first step to investigate the potential of the proposed methodology in automatically classifying rehabilitative PD speech treatment phonations as acceptable or unacceptable using a range of dysphonia measures. We have also investigated contaminating the high quality speech signals with low power noise (we used SNR=60 dB, and SNR=40 dB): specifically, we used (a) additive white Gaussian noise and (b) pink noise, which may present a more realistic setting of recording conditions at the PD subjects' homes. Encouragingly, we found practically negligible loss in performance compared to the setting where the original high quality signals were used when the noise power was kept at relatively low levels.

It would be interesting to study the effect of collecting data under less controlled settings, for example at the subjects' own homes, to verify the findings of this study: this is ultimately necessary to scale the applicability and practicality of the proposed technology. We remark that the use of the head-mounted microphone in the LSVT Companion and the expected compliance of the PD subjects to complete the speech treatment sessions in a quiet room (environment with low ambient noise) will lead with very high probability to high quality signals with minimal interference from external sources. In practice, it may be useful to integrate in the LSVT Companion an assessment of ambient noise in order to avoid starting or pausing the speech treatment session when the required acoustic requirements are not met. We have previously reported that using self-collected speech signals obtained remotely from the subjects' homes may be a viable approach towards monitoring average PD symptom severity

[14], [26]. Similarly, the LSVT Companion has shown great promise recently as a tool for independent at-home speech rehabilitation through self-collected speech signals and automatic feedback [22]. Perhaps most importantly, the satisfactory reception of the PD subjects towards the LSVT Companion makes this a promising field for further exploration [22]. We are currently working towards expanding the current database in a less acoustically controlled environment where the recordings are obtained in the subjects' homes. We will later use this information in the development of an algorithm that can predict "acceptable" and "unacceptable" ratings automatically during independent use of the LSVT Companion by PD subjects.

VI. Conclusions

We see this study as a step towards the larger goal of developing automatic decision support tools in clinical applications. We emphasize that the developed methodology is objective, fully replicable, and logistically convenient both for patients and national health systems; whereas the assessment of the phonations by trained raters is subjective, requires the dedicated time of experts, and may not always be replicable (intra-rater and inter-rater variability). The relatively limited number of samples used in the study suggests caution in the generalization of the current findings, which need to be further validated using new datasets before this technology could be used widely in clinical praxis.

We envisage these advances as major technological milestones and would set the stage for a continuum of technology-supported solutions that increase accessibility to voice clinical treatment for a wider PD population.

Acknowledgment

We want to thank all the researchers involved in the data collection process.

Declaration

We have conflict of interest to declare. L. Ramig and C. Fox have equity shares and are founders of LSVT Global, a company holding intellectual property rights on LSVT® LOUD, LSVT® BIG and LSVT® HYBRID. A. Tsanas received consulting fees from LSVT Global for this project.

References

[1] C.W. Olanow, M.B. Stern, K. Sethi, "The scientific and clinical basis for the treatment of Parkinson disease," *Neurology*, Vol. 72 (21 Suppl 4) s1-s136, 2009

[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," *European Neuropsychopharmacology*, Vol. 15, pp. 473-490, 2005

[3] A. Schrag, Y. Ben-Schlomo, N. Quinn, "How valid is the clinical diagnosis of Parkinson's disease in the community?," *Journal of Neurology, Neurosurgery Psychiatry*, Vol. 73, pp. 529-535, 2002

[4] L. Hartelius and P. Svensson, "Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: A survey," *Folia Phoniatr. Logop*., Vol. 46, pp. 9-17, 1994

[5] B. Harel, M. Cannizzaro, P.J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain and Cognition,*Vol. 56, pp. 24–29, 2004

[6] R. Pahwa, K.E. Lyons (Eds.), *Handbook of Parkinson's Disease*, 4th edition, Informa Healthcare, USA, 2007

[7] R.J. Baken and R.F. Orlikoff, *Clinical measurement of speech and voice*, San Diego: Singular Thomson Learning, 2nd ed., 2000

[8] I.R. Titze, *Principles of Voice Production*, National Center for Voice and Speech, Iowa City, US, 2nd ed., 2000

[9] V. Parsa, D.G. Jamieson, "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech," *Journal of Speech, Language, and Hearing Research*, Vol. 44, pp. 327-339, 2001

[10] M.A. Little, D.A.E. Costello, M.L. Harries, "Objective dysphonia quantification in vocal fold paralysis: comparing nonlinear with classical measures," *Journal of Voice*, Vol. 25(1), pp. 21-31, 2009

[11] A. Tsanas, P. Gómez-Vilda, "Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels," *Multidisciplinary Conference of Users of Voice, Speech and Singing* (JVHC 13), Las Palmas de Gran Canaria, pp. 3-12, 27-28 June 2013

[12] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions Biomedical Engineering*, Vol. 56 (4), pp. 1015-1022, 2009

[13] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "Accurate telemonitoring of Parkinson's disease progression using non-invasive speech tests," *IEEE Trans. Biomedical Engineering*, Vol. 57, 884-893, 2010

[14] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of the Royal Society Interface*, Vol. 8, 842-855, 2011

[15] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, Vol. 59, pp. 1264-1271, 2012

[16] J. Rusz, R. Èmejla, H. Ružièková, J. Klempíø, V. Majerová, J. Picmausová, J. Roth, E. Rùžièka, "Evaluation of speech impairment in early stages of Parkinson's disease: a prospective study with the role of pharmacotherapy," *Journal of Neural Transmission*, Vol. 120, pp. 319-329, 2013

[17] S. Skodda, W. Gronheit, U. Schlegel, "Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease," Plos One, 7(2): e32132, 2012

[18] C. Fox, L. Ramig, M. Ciucci, S. Sapir, D. McFarland, B. Farley, "The Science and Practice of LSVT/LOUD: Neural Plasticity-Principled approach to treating individuals with Parkinson disease and other neurological disorders," in *New Frontiers in Dysphagia Rehabilitation*, J. Robbins (Ed.), Seminars in Speech and Language, New York: Thieme Medical Publishers, Inc., 27, pp. 283-299, 2006

[19] L. Ramig, S. Sapir, S. Countryman, A. Pawlas, C. O'Brien, M. Hoehn, L. Thompson, "Intensive voice treatment (LSVT) for individuals with Parkinson disease: A two-year follow-up," *Journal of Neurology, Neurosurgery, and Psychiatry*, Vol. 71, pp. 493-498, 2001

[20] S. Sapir, L. Ramig, C. Fox, "Intensive Voice Treatment in Parkinson's disease: Lee Silverman Voice Treatment," *Expert Review of Neurotherapeutics*, 11, 815-830, 2011

[21] E. Taub, P.S. Lum, P. Hardin, B.W. Mark, G. Uswatte, "AutoCITE: Automated delivery of CI therapy with reduced effort by therapists," *Stroke*, 36, 1301-4, 2005

[22] A. Halpern, L. Ramig, C. Matos, J. Petska-Cable, J. Spielman, J. Pogoda, D. McFarland, "Innovative Technology for the Assisted Delivery of Intensive Voice Treatment (LSVT® LOUD) for Parkinson Disease," *American Journal of Speech-Language Pathology*, Vol. 21 (4), pp. 354-367, 2012

[23] L. Ramig, S. Countryman, L. Thompson, Y. Horii, "A comparison of two forms of intensive speech treatment for Parkinson disease," *Journal of Speech and Hearing Research*, 38, 1232-1251, 1995

[24] A. Ho, J. Bradshaw, R. Iansek, R. Alfredson, "Speech volume regulation in Parkinson's disease: effects of implicit cues and explicit instructions," *Neuropsychologia*, Vol. 37, 1453-1460, 1999

[25] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity," *International Symposium on Nonlinear Theory and its Applications* (NOLTA), pp. 457-460, Krakow, Poland, 5-8 September 2010

[26] A. Tsanas, "Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning", D.Phil. thesis, University of Oxford, Oxford, UK, 2012
[27] H. Kawahara, A. de Cheveigne, H. Banno, T. Takahashi, T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," *Interspeech*, pp. 537-540, Lisbon, Portugal, September 2005
[28] M. A. Little, P. E. McSharry, S. J. Roberts, D. Costello and I. M. Moroz," Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection," *Biomedical Engineering Online*, vol. 6 (23), 2007
[29] D. Michaelis, T. Gramss, H.W. Strube, "Glottal-to-Noise Excitation Ratio – a New Measure for Describing Pathological Voices," *Acta Acustica*, 83, 700-706, 1997
[30] P. Gomez-Vilda, J.M. Fernandez-Vicente, V. Rodellar-Biarge, R. Fernandez-Baillo, "Time-frequency representations in speech perception," *Neurocomputing*, Vol. 72, pp. 820-830, 2009
[31] S.A. Xue, D. Deliyski, "Effects of aging on selected acoustic voice parameters: preliminary normative data and educational implications," *Educational Gerontology*, 27:159–168, 2001
[32] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung, H.H. Liu, "The empirical mode decomposition and the Hilbert spectrum for non-linear and non stationary time series analysis," *Proc. Royal Soc. London A*, 454, 903-995, 1998
[33] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression," *IEEE Signal Processing Society, International Conference on Acoustics, Speech and Signal Processing* (ICASSP '10), Dallas, Texas, US, pp. 594-597, 2010
[34] J.I. Godino-Llorente, P. Gomez-Vilda, M. Blanco-Velasco, "Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters," *IEEE Transactions on Biomedical Engineering*, Vol. 53,1943-1953, 2006
[35] A. Ho, R. Iansek, C. Marigliani, J. Bradshaw, S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behavioral Neurology*, Vol. 11, 131-37, 1998
[36] A. Tsanas, M.A. Little, P.E. McSharry, "A methodology for the analysis of medical data", *in Handbook of Systems and Complexity in Health*, Eds. J.P. Sturmberg, and C.M. Martin, ISBN 978-1-4614-4997-3, Springer, New York, pp. 113-125, 2013
[37] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh (Eds.), *Feature Extraction: Foundations and Applications*, Springer, 2006
[38] C.M. Bishop, Pattern recognition and machine learning, Springer, 2007
[39] Y. Sun, S. Todorovic, S. Goodison, "Local learning based feature selection for high dimensional data analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 32, pp. 1610-1626, 2010
[40] L.Breiman, "Random Forests," *Machine Learning*, 45, 5-32, 2001
[41] V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, 1995
[42] C.C. Chang and C-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 1-27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
[43] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, Vol. 5, pp. 975-1005, 2004
[44] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer, 2nd ed., 2009
[45] N. J. Kasdin, "Discrete simulation of colored noise and stochastic processes and $1/f^{\alpha}$ power law noise generation," *Proceedings of the IEEE*, Vol.83, pp. 802-827, 1995
[46] M. Södersten, S. Ternström, M. Bohman, "Loud speech in realistic environmental noise: phonetogram data, perceptual voice quality, subjective ratings, and gender differences in healthy speakers," *Journal of Voice*, Vol. 19, pp. 29-46, 2005
[47] J. Hemphill, "Interpreting the magnitudes of correlation coefficients," *American Psychologist*, Vol. 50, pp. 78-79, 2003
[48] G.J. Meyer, S.E. Finn, L.D. Eyde, G.G. Kay, K.L. Moreland, R.R. Dies, E.J. Eisman, T.W. Kubiszyn, G.M. Reed, "Psychological testing and psychological assessment: a review of evidence and issues," *American Psychologist*, Vol. 56, pp. 128-165, 2001
[49] S. Sapir, L. Ramig, J. Spielman, C. Fox, Formant Centralization Ratio (FCR), "A proposal for a new acoustic measure of dysarthric speech," *Journal of Speech Language and Hearing Research*, 53, 114-25, 2010
[50] J.D. Arias-Londono, J.I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, G. Castellanos-Domınguez, "Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients," *IEEE Transactions on Biomedical Engineering*, Vol. 58 (2), pp. 370-379, 2011
[51] J. Rusz, J. Klempíř, E. Baborová, T. Tykalová, V. Majerová, R. Cmejla, E. Ruzicka, J. Roth, "Objective Acoustic Quantification of Phonatory Dysfunction in Huntington's Disease," *PLoS One*, 8(6): e65881, 2013 doi:10.1371/journal.pone.0065881
[52] R. Fraile, N. Saenz-Lechon, J.I. Godino-Llorente, V. Osma-Ruiz, 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

Athanasios Tsanas received the B.Sc. in Biomedical Technology Engineering from the Technological Educational Institute of Athens, Greece (2005), the B.Eng. in Electrical Engineering and Electronics from the University of Liverpool, UK (2007), the M.Sc. in Communications and Signal Processing from the Newcastle University, UK (2008) and the D.Phil. (Ph.D.) in Applied Mathematics and Biomedical Engineering from the University of Oxford, UK (2012). He is currently a Wellcome Trust post-doctoral researcher and a Junior Research Fellow of Kellogg College at the University of Oxford, UK. His research interests include signal processing and statistical machine learning, mainly applied in biomedical applications.

Max A. Little began his career writing software, signal processing algorithms and music for video games, then moved on by way of a degree in mathematics to the University of Oxford. After postdoc positions in Oxford and co-founding a web-based image search business, he won a Wellcome Trust fellowship at MIT to follow up his doctoral research in biomedical signal processing. He is currently a visiting assistant professor at MIT and an assistant professor in statistical machine learning, based at Aston University, UK.

Cynthia Fox received her doctorate degree in Speech and Hearing Sciences from the University of Arizona, Tucson. Dr. Fox is a research associate at the National Center for Voice and Speech and VP of Operations and Co-Founder of LSVT Global, Inc. She is an expert on rehabilitation and neuroplasticity and the role of exercise in the improvement of function consequent to neural injury and disease. Dr. Fox is among the world's experts in speech treatment for people with Parkinson disease. She has multiple publications in this area of focus, as well as numerous national and international research and clinical presentations.

Lorraine Ramig is a Professor in the Department of Speech-Language and Hearing Science at the University of Colorado-Boulder, a Senior Scientist at the National Center for Voice and Speech-Denver, an Adjunct Professor at Columbia University-New York City and Co-founder and President of LSVT Global, Inc. Her research has been funded by the National Institutes of Health (NIH) for over twenty years. Dr. Ramig has been a member the National Advisory Council for the National Institutes of Health-National Institutes of Deafness and Communication Disorders (NIH-NIDCD) and has received the Honors of the American-Speech-Language-Hearing Association, the highest award of her professional organization.