Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests

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

***Abstract*— Tracking Parkinson's disease (PD) symptom progression often uses the Unified Parkinson's Disease Rating Scale (UPDRS), which requires the patient's presence in clinic, and time-consuming physical examinations by trained medical staff. Thus, symptom monitoring is costly and logistically inconvenient for patient and clinical staff alike, also hindering recruitment for future large-scale clinical trials. Here, for the first time, we demonstrate rapid, remote replication of UPDRS assessment with clinically useful accuracy (about 7.5 UPDRS points difference from the clinicians' estimates), using only simple, self-administered, and non-invasive speech tests. We characterize speech with signal processing algorithms, extracting clinically useful features of average PD progression. Subsequently, we select the most parsimonious model with a robust feature selection algorithm, and statistically map the selected subset of features to UPDRS using linear and nonlinear regression techniques, which include classical least squares and non-parametric classification and regression trees (CART). We verify our findings on the largest database of PD speech in existence (~6,000 recordings from 42 PD patients, recruited to a six-month, multi-centre trial). These findings support the feasibility of frequent, remote and accurate UPDRS tracking. This technology could play a key part in telemonitoring frameworks that enable large-scale clinical trials into novel PD treatments.**

Manuscript received July 30, 2009.

Asterisk (*) indicates corresponding author.

A. Tsanas is with the Systems Analysis Modelling and Prediction (SAMP) group, University of Oxford, Oxford, UK, OX1 3LB and with the Oxford Centre for Industrial and Applied Mathematics (OCIAM), University of Oxford, Oxford, UK. (phone: 0044 1865280603; fax: 0044 1865270515; e-mail: tsanas@maths.ox.ac.uk, tsanasthanasis@gmail.com).

M. A. Little is with the SAMP group, University of Oxford, Oxford, UK, and with the OCIAM, University of Oxford, Oxford, UK (littlem@maths.ox.ac.uk).

P. E. McSharry is with the SAMP group, University of Oxford, Oxford, UK, and with the OCIAM, University of Oxford, Oxford, UK (patrick@mcsharry.net).

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).

***Index Terms*— At-Home Testing Device, classification and regression trees, speech signal processing, telemedicine, Unified Parkinson's Disease Rating Scale (UPDRS)**

I. Introduction

We are aware of neurological control through muscle movement and sensing so early in life that is easy to take it for granted. However, neurological disorders affect people profoundly and claim lives at an epidemic rate worldwide. Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's [1], and it is estimated that more than one million people in North America alone are affected [2]. Rajput *et al.* report that incidence rates have been approximately constant for the last 55 years, with 20/100,000 new cases every year [3]. A further estimated 20% of *people with Parkinson's* (PWP) are never diagnosed [4]. Moreover, these statistics are expected to increase because worldwide the population is growing older [5]. In fact, all studies suggest age is the single most important risk factor for the onset of PD, which increases steeply after age fifty [6]. Although medication and surgical intervention can hold back the progression of the disease and alleviate some of the symptoms, there is no available cure [7], [8]. Thus, early diagnosis is critical in order to improve the patient's quality of life and to prolong it [9].

The etiology of PD is largely unknown, but the symptoms result from substantial dopaminergic neuron reduction, leading to dysfunction of the basal ganglia circuitry mediating motor and some cognitive abilities [8]. *Parkinsonism* exhibits similar PD-like symptoms, but these are caused by drugs or exposure to neurotoxins for example. The main symptoms of PD are tremor, rigidity and other general movement disorders. Of particular importance to this study, vocal impairment is also common [10], [11], with studies reporting 70-90% prevalence after the onset of the disease [11]-[13]. In addition, it may be one of the earliest indicators of the disease [14], [15], and 29% of patients consider it one of their greatest hindrances [13]. There is supporting evidence of degrading performance in voice with PD progression [14], [16], [17], with *hypophonia* (reduced voice volume) and *dysphonia* (breathiness, hoarseness or creakiness in the voice) typically preceding more generalized speech disorders [11], [12].

Management of PD typically involves the administration of physical examinations applying various empirical tests, including speech and voice tests, with a medical rater *subjectively* assessing the subject's ability to perform a range of tasks. However, the necessity for the development of reliable, *objective* tools for assessing PD is manifested in the fact that current diagnosis is poor [2] and autopsy studies are reportedly inaccurate [18], [19].

Physical test observations are mapped to a metric specifically designed to follow disease progression, typically the Unified Parkinson's Disease Rating Scale (UPDRS), which reflects the presence and severity of symptoms (but does not quantify their underlying causes). For untreated patients it spans the range 0-176, with 0 representing healthy state and 176 total disability, and consists of three sections: (1) Mentation, Behavior and Mood; (2) Activities of daily living; (3) Motor. The *motor* UPDRS ranges from 0-108, with 0 denoting symptom free and 108 severe motor impairment, and encompasses tasks such as speech, facial expression, tremor and rigidity. Speech has two explicit headings, and ranges between 0-8 with 8 being unintelligible.

Noninvasive telemonitoring is an emerging option in general medical care, potentially affording reliable, cost-effective screening of PWP alleviating the burden of frequent and often inconvenient visits to the clinic. This also relieves national health systems from excessive additional workload, decreasing the cost and increasing the accuracy of clinical evaluation of the subject's condition.

The potential for telemonitoring of PD depends heavily on the design of simple tests that can be self-administered quickly and remotely. Since the recording of speech signals is non-invasive and can be readily integrated into telemedicine applications, such tests are good candidates in this regard. The use of *sustained vowel phonations* to assess the extent of vocal symptoms, where the patient is requested to hold the frequency of phonation steady for as long as possible, is common in general speech clinical practice [20] and in PD monitoring [21], [22]. This circumvents some of the confounding articulatory effects and linguistic components of *running speech* [38], i.e. the recording of standard phrases read aloud by the subject. In order to objectively characterize dysphonic symptoms, the recorded voice signals are analyzed by speech processing algorithms [22], [23].

Intel Corporation's *At-Home Testing Device* (AHTD) is a novel telemonitoring system facilitating remote, Internet-enabled measurement of a range of PD-related motor impairment symptoms, recently described in detail [24]. It records both manual dexterity and speech tests; in this study we concentrate only on sustained vowel phonations.

Previous studies have focused on separating PWP from healthy controls [14], [22]; we extend this concept to map the severity of voice symptoms to UPDRS. We also wanted to determine the feasibility of remote PD clinical trials on large scale voice data recorded in typical home acoustic environments, where previous studies have been limited to controlled acoustic environments and small numbers of recordings [22].

Recent studies have raised the important topic of finding a statistical mapping between speech properties and UPDRS as an issue worthy of further investigation, but have not addressed it explicitly [17], [24]. Here we present a method that first computes a range of classical and non-classical speech signal processing algorithms, which act as *features* for statistical regression techniques. These features establish a relationship between speech signal properties and UPDRS. We show that this method leads to clinically useful UPDRS estimation, and demonstrate remote PD monitoring on a weekly basis, tracking UPDRS fluctuations for a six-month period. This can be a useful guide for clinical staff, following the progression of clinical PD symptoms on a regular basis, tracking the UPDRS that would be obtained by a subjective clinical rater. We envisage this method finding applications in future clinical trials involving the study of large populations remote from the clinic.

II. METHODS

A. Subjects

This study makes use of the recordings described in Goetz *et al.* [24], where 52 subjects with idiopathic PD were recruited. A subject was diagnosed with PD if he had at least two of the following: rest tremor, bradykinesia (slow movement) or rigidity, without evidence of other forms of parkinsonism. The study was supervised by six US medical centers: Georgia Institute of Technology (7 subjects), National Institutes of Health (10 subjects), Oregon Health and Science University (14 subjects), Rush University Medical Center (11 subjects), Southern Illinois University (6 subjects) and University of California Los Angeles (4 subjects). All patients gave written informed consent. We disregarded data from 10 recruits – two that dropped out the study early, and a further eight that provided insufficient test data. The selected subjects had at least 20 valid study sessions during the trial period. We used data from the remaining 42 PWP (28 males) with diagnosis within the previous five years at trial onset (mean ± std. 72 ± 69, min. 1, max. 260, median 48 weeks since diagnosis), with an age range 64.4 ± 9.24, min. 36, max. 85, median 65 years. All subjects remained un-medicated for the six-month duration of the study. UPDRS was assessed at baseline (onset of trial), and after three and six months. At baseline the scores were 19.42 ± 8.12, min. 6, max. 36, median 18 points for motor UPDRS, and 26.39 ± 10.80, min. 8, max. 54, median 25.5 points for total UPDRS. After three months: 21.69 ± 9.18, min. 6, max. 38, median 21 points for motor UPDRS, and 29.36 ± 11.82, min. 7, max. 55, median 28 points for total UPDRS, and after six months: 29.57 ± 9.17, min. 5, max. 41, median 20 points for motor UPDRS, and 29.57 ± 11.92, min. 7, max. 54, median 26 points for total UPDRS.

B. Data acquisition

The data was collected using the Intel At-Home Testing Device (AHTD), which is a telemonitoring system designed to facilitate remote, Internet-enabled measurement of a variety of PD-related motor impairment symptoms. The data is collected at the patient's home, transmitted over the internet, and processed appropriately in the clinic to predict the UPDRS score. The AHTD contains a docking station for measuring tremor, paddles and pegboards for assessing upper body dexterity, a high-quality microphone headset for recording patient voice signals and a USB data stick to store test data. A LCD displays instructions for taking the tests. Typical audible prompts instruct the patient to undertake tasks to measure tremor, bradykinesia, complex co-ordinated motor function, speech and voice. As part of a trial to test the effectiveness of the AHTD system in practice, PWP were recruited and trained to use the device. Subsequently, an AHTD was installed in their home and they performed tests on a weekly basis. Each patient specified a day and time of the week during which they had to complete the protocol, prompted with an automatic alarm reminder on the device. The collected data was encrypted and transmitted to a dedicated server automatically when the USB stick was inserted in a computer with internet connection. Further details of the AHTD apparatus and trial protocol can be found in [24].

The audio recordings are of two types: sustained phonations, and running speech tests in which the subject is instructed to describe static photographs displayed on the AHTD's screen. They were recorded using a head-mounted microphone placed 5 cm from the patient's lips. The AHTD software was devised such that an initial audible, spoken instruction followed by a "beep" prompted the subject to begin phonation: an audio amplitude threshold detector triggered the capture of audio, and subsequently the capture was stopped one second after the detected signal amplitude dropped below that threshold, or 30 seconds of audio had been captured (whichever occurred sooner). The voice signals were recorded directly to the AHTD USB stick sampled at 24 KHz with 16 bit resolution.

In total, after initial screening for flawed recordings (e.g. patient coughing) where the signal was removed from the dataset, 5,923 sustained phonations of the vowel "ahhh…" were digitally processed using algorithms implemented in the Matlab software package. As explained in the introduction, we used sustained vowels to avoid the confounding effects of running speech and thereby simplify the signal analysis. The patients were required to keep their frequency of phonation as steady as possible, for as long as possible. Six phonations were recorded each day on which the test was performed: four at comfortable pitch and loudness and two at twice the initial loudness (but without shouting).

C. *Feature extraction and statistical regression techniques*

The aim of this study is to analyze the signal, extract *features* representing its characteristics, and map these features to UPDRS using regression methods. Ultimately, we want to mimic the UPDRS to useful precision with clinical importance from the speech signal. In common with other studies, we assume that vocal performance deterioration is solely due to PD and not some other pathology.

Feature extraction

Algorithms aiming to characterize clinically relevant properties from speech signals can be broadly categorized into classical *linear* and non-classical, *nonlinear* methods [22], [27]-[29]. With the term linear we refer to a method where the output is proportional to a linear combination of the inputs; conversely, *nonlinear* methods have more general relationships between the inputs and the output. Here, we applied a range of classical, and more recently proposed, speech signal processing techniques (henceforth we will collectively refer to these as '*dysphonia measures*') to all the 5,923 signals. Each of the dysphonia measures is aimed at extracting distinct characteristics of the speech signal, and produces a single number. Inevitably, some of them are highly correlated, a concept we discuss later in this paper.

The classical methods are largely based on linear signal processing techniques such as short-time autocorrelation, followed by 'peak picking' to estimate the *fundamental frequency* F0, which corresponds to the vibration frequency of the vocal folds (on average 120 Hz for men and 200 Hz for women). The pitch period is the reciprocal of F0. The voice amplitude also has clinical value and is determined as the difference between maximum and minimum values within a pitch period. Successive cycles are not exactly alike; the terms jitter and shimmer are regularly used to describe the cycle to cycle variability in F0 and amplitude, respectively. Similarly, the harmonics to noise ratio (HNR) and noise to harmonics ratio (NHR) denote the signal-to-noise estimates. We refer to references [27], [30], [36], [37] for a more detailed description of these classical speech processing techniques. The software package Praat [27] was used to calculate the classical algorithms: for comparison, the corresponding algorithms in the often-used Kay Pentax Multi-Dimensional Voice Program (MDVP) [30] are prefixed by 'MDVP' in Table 1.

The recently proposed speech signal processing methods are *Recurrence Period Density Entropy* (RPDE), *Detrended Fluctuation Analysis* (DFA) and *Pitch Period Entropy* (PPE) [22,29]. The RPDE addresses the ability of the vocal folds to sustain simple vibration, quantifying the deviations from exact periodicity. It is determined from the entropy of the distribution of the signal recurrence periods, representing the uncertainty in the measurement of the exact period in the signal. Dysphonias such as hoarseness or creaky voice typically cause an increase in RPDE. DFA characterizes the extent of turbulent noise in the speech signal, quantifying the stochastic self-similarity of the noise caused by turbulent air-flow in the vocal tract. Breathiness or other similar dysphonias caused by, e.g. incomplete vocal fold closure can increase the DFA value. Both methods have been shown to contain clinically valuable information regarding general voice disorders [29], and PD-dysphonia in particular [22]. PPE measures the impaired control of stable pitch during sustained phonation [22], a symptom common to PWP [31]. One novelty of this measure is that it uses a logarithmic pitch scale and is robust to confounding factors such as smooth vibrato which is present in healthy voices as well as dysphonic voices. It has been shown that this measure contributes significant information in separating healthy controls and PWP [22]. In total, applying the 16 dysphonia measures to the 5,923 sustained phonations, we constructed a 5,923×16 feature matrix with no invalid entries.

Data exploration and correlation analysis

In the AHTD trial, UPDRS values were obtained at baseline, three-month and six-month trial periods, but the voice recordings were obtained at weekly intervals; therefore we need to obtain weekly UPDRS estimates to associate with each phonation. The simplest approach is to use nearest neighbor interpolated UPDRS values, which would imply a sudden jump mid-way between assessments, and physiologically, this does not seem very plausible. Instead, a straightforward piecewise linear interpolation was used, with the interpolation going exactly through the measured UPDRS scores. We interpolated both motor UPDRS and total UPDRS to assess the efficacy of the dysphonia measures for predicting both scores. The tacit assumption is that symptom severity did not fluctuate wildly within the three-month intervals over which the UPDRS were obtained. Lacking actual detailed weekly UPDRS scores, linear PD progression trend is the most biologically plausible and *parsimonious* interpolation, and has been verified in a number of previous studies, many of which are reviewed in [39]. Particularly important for our argument is a recent study with non-medicated subjects followed for 12 months, which supports the use of linear UPDRS interpolation [40].

Initially, we performed correlation analysis to identify the strength of association of dysphonia measures with the linearly interpolated UPDRS values. The data was non-normal, so we used non-parametric statistical tests. We computed p-values (at the 95% level) of the null hypothesis having no correlation ρ, between each measure and UPDRS. Similarly, we calculated correlation coefficients between the dysphonia measures. We used the Spearman correlation coefficient to assess the strength of association between each measure and UPDRS, and between measures. The probability densities were computed with kernel density estimation with Gaussian kernels.

Regression mapping of dysphonia measures to UPDRS

This preliminary correlation analysis suggests that, taken individually, the dysphonia measures are weakly correlated to UPDRS. However, individual correlations alone do not reveal the (potentially nonlinear) functional relationship between these measures combined together and the associated UPDRS. To find this relationship, statistical regression techniques have been proposed, the simplest of which is classical least-squares regression [32]. Our aim is to maximally exploit the information contained in the combined dysphonia measures to produce a model that maximizes the accuracy of UPDRS prediction. We used three linear and one nonlinear regression methods to map the dysphonia measures to interpolated UPDRS values, and compared their predictive performance [32]. Linear regression methods assume that the regression function $f(\boldsymbol{x}) = y$, which maps the dysphonia measures $\boldsymbol{x} = (x_1 \dots x_M)$, where M is the number of inputs, to the UPDRS output y, is linear in the inputs. It can be expressed as $f(\boldsymbol{x}) = b_0 + \sum_{j=1}^{M} x_j b_j$, with the use of the bias term b_0 being optional, i.e. $b_0 = 0$ is quite common (this study does not use a bias term). The aim is to determine the coefficients (or parameters) $\boldsymbol{b}$, given a large number of input values $\boldsymbol{x}$ and output values $f(\boldsymbol{x}) = y$, that minimizes the error in the predictions of UPDRS over the whole data set. The linear techniques used were classical *least squares* (LS), *iteratively re-weighted least squares* (IRLS), and *least absolute shrinkage and selection operator* (Lasso). We describe these techniques next.

LS determines the coefficients $\boldsymbol{b}$ that minimize the residual sum of squares between the actual (measured) UPDRS and the predicted UPDRS:

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} \left(y_i - f(\boldsymbol{x}_i)\right)^2 = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{M} x_{ij} b_j \right)^2$$

where $\boldsymbol{x}_i = (x_{i1} \dots x_{iM})$ is a vector of input measurements giving rise to the measured quantity y_i, for each ith case and N is the number of observations. The statistical assumption underlying LS is that the residuals (the difference between the actual and predicted UPDRS) are independent and identically distributed Gaussian random variables, which may not always be a valid assertion, and this can lead to poor estimates of the parameters. Thus, to mitigate any large deviations from Gaussianity, our proposed IRLS method effectively reduces the influence of values distant from the main bulk of the data (*outliers*) by making iterative LS predictions that reweight outliers at each step. This robust estimator is computed using the following algorithm:

1) Determine the residuals: $\boldsymbol{r} = \sum_{i=1}^{N} \left| y_i - \sum_{j=1}^{M} x_{ij} b_j \right|$
2) Determine the weights w using r: $\boldsymbol{w} = \exp[-2\boldsymbol{r}/\max(\boldsymbol{r})]^T$
3) Solve the least squares problem using $\mathbf{w}$:
$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} w_i \left(y_i - \sum_{j=1}^{M} x_{ij} b_j\right)^2$
4) Repeat from the first step, for a pre-specified number of iterations (we used 100). In the first iteration, the coefficients $\boldsymbol{b}$ are determined using the LS method.

A problem often encountered in such regression methods when using a large number of input variables (16 in this case) is the *curse of dimensionality*: fewer input variables could potentially lead to a simpler model with more accurate prediction. Research has shown that many of the dysphonia measures are highly correlated [22] and this finding is confirmed in this study (see Table 2), so we can assume that taken together, highly correlated measures contribute little additional information for UPDRS prediction. Following the general *principle of parsimony*, we would like to reduce the number of measures in the analysis and still obtain accurate UPDRS prediction.

The Lasso is a principled *shrinkage method* that has, relatively recently, emerged as a powerful feature selection tool, which also offers a mathematical framework enhancing the physiological interpretability of the resulting regression coefficients [33]. The Lasso has the desirable characteristic of simultaneously minimizing the prediction error whilst producing some coefficients that are effectively zero (reducing the number of relevant input variables) by adjusting a *shrinkage* parameter. The algorithm selects the best, smallest subset of variables for the given shrinkage parameter. Decreasing this parameter value causes additional coefficients to shrink towards zero, further reducing the number of relevant input variables. Then it becomes a matter of experimentation to find the optimal compromise between reducing the number of relevant input measures and minimizing the error in the UPDRS prediction. Specifically, the Lasso induces the *sum of absolute values penalty*: $\hat{\boldsymbol{b}}_{LASSO} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{M} x_{ij} b_j\right)^2$ subject to $\sum_{j=1}^{M}\left|b_j\right| \le t$, where t is the shrinkage parameter, and the constraint $\sum_{j=1}^{M}\left|b_j\right| \le t$ can be seen as imposing the penalty $\lambda\sum_{j=1}^{M}\left|b_j\right|$ to the residual sum of squares, which yields:

$$\hat{\boldsymbol{b}}_{LASSO} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{M} x_{ij} b_j\right)^2 + \lambda\sum_{j=1}^{M}\left|b_j\right|$$

Other penalties are possible, including the sum of squares of coefficients $\boldsymbol{b}$ (ridge regression), but it can be shown that the sum of absolute values penalty leads to many coefficients which are almost exactly zero, when the problem is underdetermined due to highly redundant inputs, as in this case [26]. In practical terms, this also enhances the *interpretability* of the model.

It may well be the case that the dysphonia measures do not combine linearly to predict the UPDRS. Thus, *nonlinear regression* may be required, where the prediction function $f(\boldsymbol{x})$ is a nonlinear combination of the inputs $\boldsymbol{x}$. To test this idea, we used the *classification and regression tree* (CART) method, which is a conceptually simple nonlinear method that often provides excellent regression results [32]. The key idea behind CART is in finding the best split of the input variables, and partitioning the ranges of these variables into two sub-regions. This partitioning process is repeated on each of the resulting sub-regions, recursively partitioning the input variables into smaller and smaller sub-regions. This recursive procedure can be represented graphically as a tree that splits into successively smaller branches, each branch representing a sub-region of input variable ranges. This tree is "grown" up to T_0 splits, learning a successively detailed mapping between all the available data and the UPDRS. Although this process is in principle very flexible and hence able to reproduce highly convoluted mappings, it can easily *overfit* the data: that is, become highly sensitive to noisy fluctuations in the input data. To address this danger some splits are collapsed (a process known as *pruning*) and the amount of split reduction is determined by the *pruning level*.

Here we employed the following strategy: we have experimented with the Lasso method by adjusting the constant parameter λ, and then observed the surviving and shrinking coefficients associated with each dysphonia measure. Subsequently, various reduced sets of dysphonia measures have been tested with all the regression methods (LS, IRLS and CART).

Model selection – Bayesian Information Criterion and Akaike Information Criterion

The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) offer a framework of comparing fits of models with a different number of parameters [32], and have often been used in the context of medical applications [34]. These criteria induce a penalty on the number of measures in the selected subset, offering a compromise between in-sample error and model complexity. The 'optimal' subset of dysphonia measures is the model with the lowest BIC and AIC values. Assuming the errors are Gaussian, these two criteria are defined as [32], $BIC = \sum_i^N\left(U_i - \hat{U}_i\right)^2/\sigma_\varepsilon^2 + \log(N)D$, $AIC = \sum_i^N\left(U_i - \hat{U}_i\right)^2/\sigma_\varepsilon^2 + 2D$, where N is the number of data samples, D is the number of measures, U_i is the true UPDRS value as provided by the dataset, $\hat{U}_i$ the predicted estimate and σ_ε^2 is the mean squared error (MSE) variance, where $MSE = \frac{1}{N}\sum_i^N\left(U_i - \hat{U}_i\right)^2$.

D. *Cross-validation and model generalization*

To objectively test the generalization performance of the proposed regression methods in predicting UPDRS (that is, the ability of the models to perform well on data not used in estimating the model parameters), we used cross validation, a well-known statistical re-sampling technique [35]. Specifically, the data set of 5,923 phonations was split into a *training* subset (5,331 phonations) and a *testing* subset (592 phonations), which was used to assess generalization performance. The model parameters were derived using the *training* subset, and errors were computed using the *testing* subset (out-of-sample error or testing error). The process was repeated a total of 1,000 times, with the data set randomly permuted in each run prior to splitting into training and testing subsets, in order to obtain confidence in this assessment. On each test repetition, we recorded the *mean absolute error* (MAE) for both training and testing subsets $MAE = \frac{1}{N}\sum_{i\in Q}\left|U_i - \hat{U}_i\right|$, where N is the number of phonations in the training or testing dataset, denoted by Q, containing the indices of that set. Testing errors from all 1,000 repetitions were averaged. In all cases, the prediction performance results were determined following cross-validation.

III. RESULTS

A. *Data exploration and correlation analysis*

Speech appears explicitly in sections 5 and 18 of the UPDRS metric. These entries, taken together, are strongly correlated to motor-UPDRS (p<0.001, Spearman ρ=0.44) and total-UPDRS (p<0.001, ρ=0.51), indicating strong association between speech and UPDRS. These statistically significant findings intuitively suggest that the extraction of subtle features from speech signals could accentuate this concealed relationship.

Table 1 summarizes the dysphonia measures used in this study. All measures were significantly correlated (p<0.001) with linearly interpolated motor-UPDRS and total-UPDRS scores. Although statistically significant, none of the measures taken individually appears to have a large magnitude of correlation to either motor or total-UPDRS. Following normalization to the range 0 to 1, the probability densities of each dysphonia measure are shown in Fig. 1a. The jitter,

TABLE I
CLASSICAL AND NON-CLASSICAL DYSPHONIA MEASURES APPLIED TO SUSTAINED VOWEL PHONATIONS, AND THEIR UPDRS CORRELATIONS.

Measure	Description	Motor UPDRS correlation	Total UPDRS correlation
MDVP: Jitter(%)	KP-MDVP jitter as a percentage	0.124	0.125
MDVP: Jitter(Abs)	KP-MDVP absolute jitter in microseconds	0.072	0.103
MDVP:RAP	KP-MDVP Relative Amplitude Perturbation	0.105	0.107
MDVP:PPQ	KP-MDVP five-point Period Perturbation Quotient	0.120	0.117
Jitter:DDP	Average absolute difference of differences between cycles, divided by the average period	0.105	0.107
MDVP: Shimmer	KP-MDVP local shimmer	0.138	0.139
MDVP: Shimmer(dB)	KP-MDVP local shimmer in decibels	0.139	0.139
Shimmer: APQ3	Three point Amplitude Perturbation Quotient	0.116	0.122
Shimmer: APQ5	Five point Amplitude Perturbation Quotient	0.123	0.127
MDVP:APQ	KP-MDVP 11-point Amplitude Perturbation Quotient	0.166	0.163
Shimmer: DDA	Average absolute difference between consecutive differences between the amplitudes of consecutive periods	0.116	0.122
NHR	Noise-to-Harmonics Ratio	0.131	0.139
HNR	Harmonics-to-Noise Ratio	-0.159	-0.163
RPDE	Recurrence Period Density Entropy	0.112	0.143
DFA	Detrended Fluctuation Analysis	-0.131	-0.141
PPE	Pitch Period Entropy	0.160	0.152

KP-MDVP stands for Kay Pentax Multidimensional Voice Program. Classical measures were obtained using the Praat software package. The UPDRS correlation columns are the Spearman non-parametric correlation coefficient between each measure and piecewise linearly interpolated motor and total UPDRS. All measures were statistically significantly correlated ($p < 0.0001$) with motor-UPDRS and total-UPDRS. All speech signals were used to generate these results (n = 5,923 phonations).

shimmer and NHR measures are distributed close to zero, whereas HNR, RPDE, DFA and PPE are more evenly distributed. Table 2 presents the Spearman rank-correlations between all the dysphonia measures. All measures were statistically significantly correlated (p<0.001). Fig. 1 (b, c) displays the normalized dysphonia measures against motor and total-UPDRS, providing an indication of their associated relationship to UPDRS.

B. Regression analysis

Table 3 presents the regression coefficient values for all dysphonia measures, for all three linear prediction methods. The obtained coefficients differed over cross-validation runs for all three linear models, as evidenced by the large standard deviation of some of the coefficients. However, the testing mean absolute error (MAE) and its standard deviation across the 1,000-run cross-validation was relatively low, suggesting that these indicative coefficients are sufficient for useful

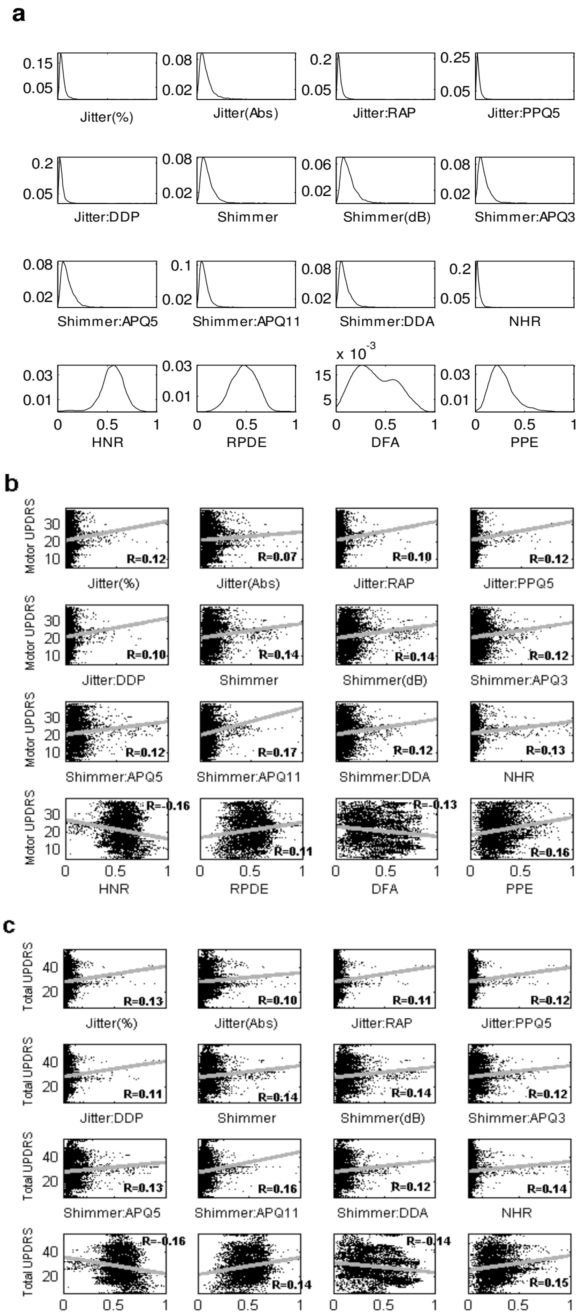

Fig. 1. a) Probability densities of the dysphonia measures applied to the 5,923 sustained phonations. The vertical axes are the probability densities of the normalized measures, estimated using kernel density estimation with Gaussian kernels. b) Dysphonia measures against motor UPDRS, and c) Dysphonia measures against total UPDRS. The horizontal axes are the normalized dysphonia measures and the vertical axes correspond to UPDRS. The grey lines are the best linear fit obtained using IRLS - see methods section for description of the algorithm. The R-values denote the Spearman correlation coefficient of each measure with UPDRS.

TABLE II

CORRELATION COEFFICIENTS BETWEEN DYSPHONIA MEASURES.

	MDVP: Jitter (%)	MDVP: Jitter (Abs)	MDVP:RAP	MDVP: PPQ	Jitter: DDP	MDVP: Shimmer	MDVP: Shimmer (dB)	Shimmer APQ3	Shimmer APQ5	MDVP: APQ	Shimmer DDA	NHR	HNR	RPDE	DFA
MDVP: Jitter(Abs)	0.90														
MDVP: RAP	***0.96***	0.82													
MDVP:PPQ	***0.96***	0.89	***0.95***												
Jitter:DDP	***0.96***	0.82	***1***	***0.95***											
MDVP:Shimmer	0.65	0.63	0.65	0.69	0.65										
MDVP: Shimmer(dB)	0.68	0.64	0.66	0.70	0.66	***0.99***									
Shimmer: APQ3	0.62	0.58	0.63	0.66	0.63	***0.98***	***0.96***								
Shimmer: APQ5	0.62	0.61	0.62	0.67	0.62	***0.99***	***0.97***	***0.98***							
MDVP:APQ	0.63	0.64	0.60	0.67	0.60	***0.96***	***0.95***	0.91	***0.96***						
Shimmer:DDA	0.62	0.58	0.63	0.66	0.63	***0.98***	***0.96***	***1***	***0.98***	0.91					
NHR	0.80	0.75	0.75	0.75	0.75	0.65	0.69	0.62	0.62	0.62	0.62				
HNR	-0.76	-0.76	-0.73	-0.79	-0.73	-0.80	-0.78	-0.78	-0.79	-0.79	-0.78	-0.76			
RPDE	0.53	0.64	0.45	0.51	0.45	0.48	0.47	0.43	0.46	0.50	0.43	0.61	-0.65		
DFA	0.44	0.50	0.43	0.48	0.43	0.29	0.27	0.26	0.29	0.31	0.26	0.15	-0.36	0.19	
PPE	0.85	0.81	0.77	0.84	0.77	0.64	0.66	0.59	0.62	0.66	0.59	0.73	-0.75	0.55	0.42

The correlation columns are the Spearman non-parametric correlation coefficients ρ between two measures. All measures were statistically significantly correlated ($p < 0.0001$). Bold italic entries indicate high correlation between measures (Spearman $\rho \geq 0.95$). All speech signals were used (n = 5,923).

TABLE III

REGRESSION COEFFICIENTS OF LS, IRLS, AND LASSO FOR ALL DYSPHONIA MEASURES AND PIECEWISE LINEARLY INTERPOLATED MOTOR AND TOTAL UPDRS.

Measure	Motor UPDRS LS coefficients	Motor UPDRS IRLS coefficients	Motor UPDRS Lasso coefficients (λ=1)	Total UPDRS LS coefficients	Total UPDRS IRLS coefficients	Total UPDRS Lasso coefficients (λ=1)
MDVP:Jitter (%)	-87.63	-183.28	-214.45	-768.96	-649.19	-537.90
MDVP:Jitter(Abs)	$-6.87\cdot10^{4}$	$-7.64\cdot10^{4}$	0	$-7.04\cdot10^{4}$	$-8.49\cdot10^{4}$	0
MDVP: RAP	$-6.02\cdot10^{4}$	$-6.29\cdot10^{4}$	0	$-2.91\cdot10^{4}$	$-3.36\cdot10^{4}$	0
MDVP: PPQ	-238.07	-62.70	0	209.26	40.02	50.62
Jitter:DDP	$2.02\cdot10^{4}$	$2.12\cdot10^{4}$	75.59	$1.02\cdot10^{4}$	$1.17\cdot10^{4}$	241.81
MDVP:Shimmer	77.78	100.56	23.81	28.62	114.26	9.58
MDVP:Shimmer (dB)	0.31	-2.49	4.37	-0.38	-4.74	1.67
Shimmer:APQ3	$-1.85\cdot10^{4}$	$-2.43\cdot10^{4}$	0	$-8.19\cdot10^{4}$	$-7.24\cdot10^{4}$	0
Shimmer:APQ5	-108.01	-126.06	-66.68	-93.05	-138.32	-2.75
MDVP:APQ	55.12	83.35	66.28	104.35	107.95	85.74
Shimmer:DDA	$6.16\cdot10^{3}$	$8.09\cdot10^{3}$	-4.97	$2.73\cdot10^{4}$	$2.41\cdot10^{4}$	0
NHR	2.14	-5.04	-7.38	-12.45	-8.21	-17.33
HNR	0.52	0.57	0.61	0.65	0.74	0.74
RPDE	16.62	20.24	15.25	26.21	30.77	23.81
DFA	-9.54	-15.43	-12.05	-12.47	-19.73	-14.05
PPE	35.34	37.90	28.50	41.37	39.15	33.41

The coefficients in this table are indicative (derived over one run of cross-validation with the training subset, n = 5,331). We have noticed considerably different values in the 1,000 runs of 10-fold cross validation. However, the fact that the cross-validated test error and test error standard deviation remained small, suggests that confidence can be assumed for the above coefficient values.

UPDRS prediction. The training MAE for the linearly interpolated motor-UPDRS was 6.7 for LS and IRLS, and 6.8 for Lasso. The testing MAE was 6.7 for LS and IRLS, and 6.8 for Lasso. The CART method outperforms the linear predictors with a training MAE of 4.5 and testing MAE of 5.8. The training error for the linearly interpolated total-UPDRS was 8.5 for LS, 8.4 for IRLS, and 8.5 for Lasso. The testing error was 8.5 for LS, 8.4 for IRLS, and 8.6 for Lasso. CART performs better again, producing a training MAE of 6.0 and testing MAE of 7.5. IRLS is slightly superior compared to the linear predictors. However, CART outperforms it, displaying the smallest deviation from the interpolated score.

C. *Model selection and validation*

Sweeping the Lasso algorithm regularization parameter λ, we derived a number of dysphonia subsets. The pruning level for CART was set to minimize the MAE, following manual spot-checks. We noted that a difference in value of up to 20 for the pruning level did not produce significantly different results, given that the number of splits of the data was large. Both information criteria (AIC and BIC) agree on a subset containing six measures: MDVP: Jitter (Abs), MDVP: Shimmer, NHR, HNR, DFA, PPE for the CART method. This subset gives testing errors of 6.80±0.17 for motor-UPDRS and 8.47±0.27 for total UPDRS using the IRLS method. Similarly, we obtain 5.95±0.19 and 7.52±0.25 using the CART method. The testing errors remain low and close to the training error, indicating that the model has achieved a reasonable estimate of the performance we might expect on novel data. The difference between predicted and linearly interpolated UPDRS values is typically low.

In Fig. 2 we demonstrate UPDRS tracking of the patient with the severest fluctuation throughout the six-month trial for the best linear method, IRLS, and for CART. CART achieves

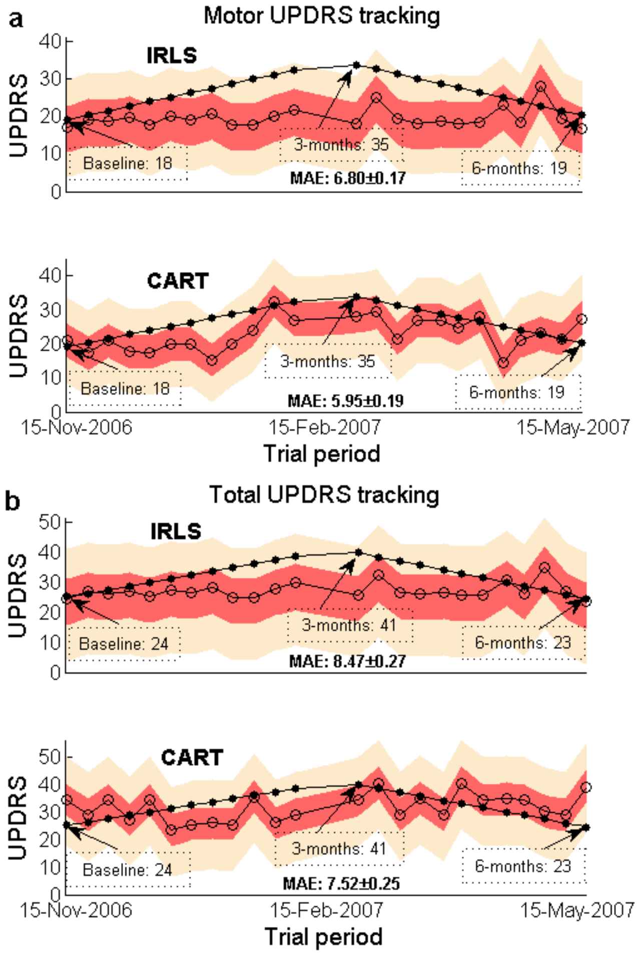

Fig. 2. a) Motor UPDRS and b) total-UPDRS tracking over the 6-month trial period for the patient with the severest fluctuation (sharp UPDRS increase mid-way and subsequent decrease). The 'baseline', '3-month' and '6-month' UPDRS scores are shown. The dots denote the piecewise linearly interpolated UPDRS value and the circles, predicted UPDRS. The light brown bands are the 5-95 percentile confidence interval of the UPDRS prediction, and the red bands are the 25-75 percentile confidence intervals. Confidence intervals are estimated using 1,000-runs of 10-fold cross-validated out-of-sample UPDRS prediction. The MAE of each model is also quoted, along with the standard deviation. The CART method tracks Parkinson's disease symptom progression more accurately than IRLS. The out-of-sample MAE was computed by taking the average MAE of the 1,000 runs of the cross-validation of each testing subset (n = 592 phonations).

the smallest prediction error, and tracks the linearly interpolated UPDRS more accurately. We have purposefully chosen a patient with a highly irregular UPDRS pattern (sharp increase and mid-way subsequent decrease) to demonstrate the performance of the regression algorithms at tracking UPDRS. In results not shown here, we note that patients with smaller UPDRS changes over the six-month trial can be monitored with even greater accuracy. In general, PD leads to increasing UPDRS scores in the long run. However, these scores may not be monotonically increasing over shorter intervals for all patients (such as in the subject presented in Fig. 2).

IV. DISCUSSION

In this study, we have examined the potential of sustained vowel phonations in predicting average PD symptom progression, establishing a mapping between dysphonia measures and UPDRS. The association strength of these measures and (motor and total) UPDRS was explored, using three linear and one nonlinear regression method. We have selected an optimally reduced subset of the measures producing a clinically useful model, where each measure in the subset extracts non-overlapping physiological characteristics of the speech signal. The comparatively small MAE is notable: the sustained vowel phonations convey sufficient information to predict UPDRS to clinically useful accuracy. It has been demonstrated that motor-UPDRS can be estimated within approximately 6 points (the full range spans 108 points) and total-UPDRS within 7.5 points (the full range spans 176 points), predictions which are very close to the clinicians' observations. These results reflect the best estimate of the *asymptotic out-of-sample* prediction error using the 1,000 runs 10-fold cross-validation scheme. It is true that the nearly 6,000 samples come from 42 patients which could lead to some dependence between the samples, dependence that might affect the reliability of the cross-validation. However, only a small number of patients were recruited to the study, and any patient-specific cross-validation is not really reliable: there is not enough hold-out data and in our own experimental computations the standard deviation of the errors was too large. Therefore, simple patient-specific cross-validation is too unstable to form a reliable estimate of the asymptotic out-of-sample prediction error.

Furthermore, we showed the feasibility of tracking UPDRS changes in time (Fig. 2). Perhaps most importantly, the satisfactory reception of the patients themselves towards the AHTD and speech tests [24] makes this a promising field for further experimentation. The 42 PWP in the present study were diagnosed within the previous five years at trial onset and displayed moderate symptoms (max motor-UPDRS 41, max total-UPDRS 55), so it would be important to look at a more severely impaired group in the future. The satisfactory UPDRS estimation in moderate symptoms, which are difficult to detect, accentuates the potential of the dysphonia measures in PD assessment and supports the feasibility of successful UPDRS tracking in more severely affected patients. It is conceivable that PWP at later stages could be monitored at least as accurately, due to their more pronounced vocal symptoms.

Speech appears explicitly in two UPDRS categories (part II, activities of daily living section and part III, motor section). One could argue that speech is more strongly related to the motor section rather than daily living activities and mentation, behavior and mood (part I), because the underlying etiology of dysphonic sustained phonations may be physiologically attributed to flawed muscle control, most likely caused by dopaminergic neuron reduction. This would imply that *only* motor-UPDRS estimation would be tractable. However, the results of this study indicate that total-UPDRS estimation with clinically useful accuracy is plausible, suggesting that PD speech dysphonias could be at least partly related to mood as well. This makes it possible to suggest the generalization that the underlying causes of PD symptoms such as tremor and mood are manifested in impaired speech control. We can only speculate on the underlying biological causes, but the correlation coefficients reveal statistically significant and strong associations between speech and motor function

(ρ=0.44), and between speech and general health deterioration, including mood (ρ=0.51). Stebbins *et al.* [25] reported that motor-UPDRS can be explained by six distinct and clinically useful, underlying factors: speech, facial expression, balance and gait (factor I), rest tremor (factor II), rigidity (factor IV), right and left bradykinesia (factors III and V), and postural tremor (factor VI). They found relatively low correlations between the six factors, suggesting all contribute to accurate UPDRS estimation by capturing different aspects of PD symptoms. In terms of that study, we have used measures *within* factor I, extracting PD information properties only from speech. The implicit argument is that the dysphonia measures can adequately reveal PD symptom severity estimated by UPDRS, because they capture the effects of PD motor impairment manifested in speech production. We have demonstrated that predicting both motor and total-UPDRS scores to useful precision is possible, because the dysphonia measures aid in uncovering functional features of PD impairment.

Additionally, our findings support the argument that non-classical dysphonia measures convey important information for clinical speech signal processing. This is evidenced in the results of the Lasso algorithm, which selected non-standard dysphonia measures in all the performed tests (especially HNR, RPDE, DFA and PPE), and reflected in the optimal dysphonia measure subset selected by the AIC and BIC. This suggests that these dysphonia measures contain significant information for tracking UPDRS. It also reinforces the conclusion reached in a previous study [22], where these non-standard measures outperformed their classical counterparts in separating PWP from healthy controls. Nevertheless, the classical measures convey useful information which may not be captured by the non-classical techniques: a parsimonious combination of classical and non-classical is optimal. That is, different dysphonia measures appear to characterize different aspects of the PD symptoms represented in the speech signal, so that their combination in a regression method captures properties useful for clinical purposes.

In a general statistical regression setting, some variables (here the dysphonia measures) will be mapped to a target variable (here UPDRS). *Linear* regression is a simple and often adequate approach, hence providing a benchmark against which more complicated *nonlinear* regression methods can be compared. Interestingly, the linear predictors used in this study performed very well, with the IRLS always presenting slightly better prediction results than LS and Lasso. This indicates that the tails of the error distributions of UPDRS around the regression line may depart from Gaussianity and outliers need to be eliminated from the Gaussian prediction error supposed by classical least squares methods. Still, its performance is not usefully superior to the standard linear LS method. However, the linear regression line may not be a good model, which often suggests the use of *nonlinear* predictors. We used CART, which is acknowledged as the best “off-the-shelf” method for predictive learning [32]. CART always provided approximately 1-2 UPDRS points' improvement in prediction performance over the linear methods.

Some of the dysphonia measures are highly correlated with each other (Table 2), which suggested the removal of those with insignificant contribution towards UPDRS estimation. This large correlation between measures manifests in the parameter values obtained through LS regression, where two highly correlated measures are allocated opposite signed, but similar magnitude, large value parameters. For example, the measures Shimmer APQ5 and MDVP: APQ have a correlation coefficient 0.96 and their parameters almost exactly cancel each other. To address this artifact, the Lasso algorithm offers a principled mathematical framework for reducing the number of relevant input variables. Furthermore, recent theoretical work has shown that, remarkably, where there is a subset of input measures that contribute no additional information over others in the set, this algorithm is essentially *equivalent to a brute force search* through all possible combinations of measures to find the smallest combination that produces the minimum prediction error [26].

The principle of parsimony suggests that given several different combinations of dysphonia measures that have equal prediction accuracy, preference should be given to the combination with the smallest number of measures. To account for estimation precision versus model complexity (number of dysphonia measures in the subset), we used the AIC and BIC values to determine the ‘optimal’ subset. Both criteria suggest using the subset with the six measures: (MDVP: Jitter (Abs), MDVP: Shimmer, NHR, HNR, DFA, PPE) in combination with the CART method, which offers an attractive compromise between performance and complexity. That is, the selected dysphonia measures in this subset complement each other with minimal overlapping information, and at the same time capture practically the entire range of possible differentiating features of the speech signals useful in determining UPDRS values.

This selected subset and associated coefficients can be given a tentative physiological interpretation. Fundamental frequency variations (measured with absolute jitter) and variations in signal amplitude (shimmer), are well established methods, capturing symptoms manifested in vocal fold vibration and lung efficiency. NHR and HNR suggest that UPDRS is affected by increased noise, caused by turbulent airflow in the glottis, often resulting from incomplete closure of the vocal folds. This concept is further backed up by the inclusion of DFA. Finally PPE indicates impaired pitch control which could be interpreted as deteriorating muscle co-ordination. This is a sign of flawed neuron action potential averaging, suggesting the reduction of dopaminergic neurons devoted to speech control. The remaining dysphonia measures were shown to convey insignificant *additional* information to be included in the model.

We believe these exploratory results could be of value in clinical trials, presenting clinical staff with a useful guide to clinical rater tracking of PD symptoms by UPDRS remotely, and at weekly intervals. This could be particularly useful in those cases where the patients are reluctant or unable to make frequent physical visits to the clinic. This may also be invaluable for future clinical trials of novel treatments which will require high-frequency, remote, and very large study populations. We remark that it is highly likely that combining these results with other PD symptom measures such as those obtained using the AHTD dexterity tests may well help to reduce the UPDRS prediction error and enhance the clinical value of such multimodal testing in telemedicine applications.

We stress again the fact that UPDRS is *subjective*, and the clinicians' verdict on a patient's score could vary. In the end, often the most relevant aspect of disease progression (or PD treatment) is the patient's *perception* of symptoms, i.e. symptom self-rating. This study was confined to using dysphonia measures to predict the average clinical overview of the widely used PD metric, the UPDRS. Although the dysphonia measures have physiological interpretations, it is difficult to link self-perception and physiology. In ongoing research work we focus our attempts to establish a more physiologically-based model, which will explain the data-driven findings in this study in terms of the relevant physiological changes that occur in PD.

ACKNOWLEDGMENT

We are grateful to Ralph Gregory for medical insight and to Mike Deisher, Bill DeLeeuw and Sangita Sharma at Intel Corporation for fruitful discussions and comments on early drafts of the paper. We also want to thank James McNames, Lucia M. Blasucci, Eric Dishman, Rodger Elble, Christopher G. Goetz, Andy S. Grove, Mark Hallett, Peter H. Kraus, Ken Kubota, John Nutt, Terence Sanger, Kapil D. Sethi, Ejaz A. Shamim, Helen Bronte-Stewart, Jennifer Spielman, Barr C. Taylor, David Wolff, and Allan D. Wu, who were responsible for the design and construction of the AHTD device and organizing the trials in which the data used in this study was collected.

REFERENCES

[1] M.C. de Rijk et al. Prevalence of Parkinson's disease in Europe: a collaborative study of population-based cohorts. *Neurology*, Vol. 54, pp. 21–23, 2000

[2] A.E. Lang, A.M. Lozano. Parkinson's disease – First of two parts, *New England Journal Medicine*, 339, 1044-1053, 1998

[3] M. Rajput, A. Rajput, A.H. Rajput. Epidemiology (chapter 2). In *Handbook of Parkinson's disease*, edited by R. Pahwa and K. E. Lyons, 4th edition, Informa Healthcare, USA, 2007

[4] 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

[5] S.K. Van Den Eeden et al.. Incidence of Parkinson's disease: Variation by age, gender, and Race/Ethnicity. *Am J Epidem* 157, 1015-1022, 2003

[6] A. Elbaz *et al*. Risk tables for parkinsonism and Parkinson's disease. *Journal of Clinical Epidemiology*. Vol. 55, pp. 25-31, 2002

[7] S. Sapir, J. Spielman, L. Ramig, B. Story, C. Fox. Effects of Intensive Voice Treatment (LSVT) on Vowel Articulation in Dysarthric Individuals with Idiopathic Parkinson Disease: Acoustic and Perceptual Findings. *J. Speech, Lang. Hear Research* , Vol.50, pp. 899-912, 2007

[8] N. Singh, V. Pillay, Y. E. Choonara. Advances in the treatment of Parkinson's disease, *Progress in Neurobiology*, Vol. 81, pp. 29-44, 2007

[9] J. King, L. Ramig, J. H. Lemke, Y. Horii. Parkinson's disease: longitudinal changes in acoustic parameters of phonation, *Journal of Medical Speech and Language Pathology* 2, 29-42 (1994)

[10] D. Hanson, B. Gerratt and P. Ward. Cinegraphic observations of laryngeal function in Parkinson's disease. *Laryngoscope* 94, 348-353, 1984

[11] A. Ho, R. Iansek, C. Marigliani, J. Bradshaw, S. Gates. Speech impairment in a large sample of patients with Parkinson's disease. *Behavioral Neurology* 11, 131-37, 1998

[12] J.A. Logemann, H.B. Fisher, B. Boshes, E. R. Blonsky. Frequency and coocurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *J. Speech Hear. Disord.* 43, 47-57, 1978

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

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

[15] J.R. Duffy. *Motor Speech Disorders: substrates, differential diagnosis and management*, New York: Mosby, 2nd ed., 2005

[16] R.J. Holmes, J.M. Oates, D.J. Phyland, A.J. Hughes. Voice characteristics in the progression of Parkinson's disease. *Int J Lang Comm Dis* 35, 407-418, 2000

[17] S. Skodda, H. Rinsche, U. Schlegel. Progression of dysprosody in Parkinson's disease over time – A longitudinal study. *Movement Disorders* 24 (5), 716-722, 2009

[18] A. H. Rajput, B. Rozdilsky, A. Rajput. Accuracy of clinical diagnosis in parkinsonism – a prospective study, *Canadian Journal of Neurological Sciences* 18 (3), 275-278, 1991

[19] A.J. Hughes, S.E. Daniel, S. Blankson, A.J. Lees. A clinicopathologic study of 100 cases of Parkinson's disease, *Archives of Neurology* 50, 140–148, 1993

[20] I.R. Titze. Summary statement: Workshop on Acoustic Voice Analysis, (available online at: http://www.ncvs.org/museum-archive/sumstat.pdf, last accessed on 9 Sep. 2009) NCVS, Denver, Colorado, Feb. 1994

[21] K.M. Rosen, R.D. Kent, J.R. Duffy. Task-based profile of vocal intensity decline in Parkinson's disease. *Folia Phoniatr. Logop* 57, 28-37, 2005

[22] 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 Trans. Biomedical Engineering* 56(4), 1015-1022, 2009

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

[24] C.G. Goetz. *et al.* Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device. *Movement Disorders* 24 (4), 551-556, 2009

[25] G. T. Stebbins, C.G. Goetz, A.E. Lang, E. Cubo. Factor analysis of the motor section of the Unified Parkinson's Disease Rating Scale during the off-state. *Movement Disorders* 14 (4), 585-589, 1999

[26] D. Donoho. For most large underdetermined systems of equations, the minimal L1-norm near-solution approximates the sparsest near-solution. *Commun. Pure and Applied Mathematics*. 59(7), 904-934, 2006

[27] P. Boersma and D. Weenink: Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/, 2009

[28] M.A. Little. *Biomechanically Informed Nonlinear Speech Signal Processing*, DPhil Thesis, University of Oxford, Oxford, UK, 2007

[29] M.A. Little, P.E. McSharry, S.J. Roberts, D. Costello, I.M. Moroz. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *Biomedical Engineering Online* 6:23, 2007

[30] KayPENTAX, Kay Elemetrics Disordered Voice Database, Model 4337, Kay Elemetrics, Lincoln Park, NJ, USA, 1996-2005

[31] L. Cnockaert *et al.* Low frequency vocal modulations in vowels produced by Parkinsonian subjects, *Speech Comm* 50, 288-300, 2008

[32] T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd ed., 2009

[33] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. J. *R. Statist. Soc. B* 58, 267-288, 1996

[34] N. Stergiopulos, B.E. Westerhof, N. Westerhof. Total Arterial Inertance as the fourth element of the windkessel model. *Am. J. Physiol. Heart Circ. Physiol.* 276, 81-88, 1999

[35] A. Webb. *Statistical Pattern Recognition*, John Wiley and Sons Ltd, 2002

[36] 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*, (*in press)*

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

[38] J. Schoentgen and R. De Gucteneere: Time series analysis of jitter, *Journal of Phonetics*, Vol. 23, pp. 189-201, 1995

[39] P.L.S. Chan and N.H.G. Holford: Drug treatment effects on disease progression, *Annual Review of Pharmacology and Toxicology*, Vol. 41, pp. 625-659, 2001

[40] M.W.M. Schüpbach, J-C. Corvol, V. Czernecki, M.B. Djebara, J-L. Golmard, Y. Agid and A. Hartmann: The segmental progression of early untreated Parkinson disease: a novel approach to clinical rating, *J. Neurol. Neurosurg. Psychiatry*, doi:10.1136/jnnp.2008.159699