Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity

Athanasios Tsanas[a,b,*], Max A. Little[a,b,c], Patrick E. McSharry[a,b,d], Lorraine O. Ramig[e,f]

* *Asterisk denotes corresponding author.* (Tel. 0044 1865 280603)

[a] Systems Analysis, Modelling and Prediction (SAMP) group, Mathematical Institute and Department of Engineering Science, University of Oxford, Oxford, UK

[b] Oxford Centre for Industrial and Applied Mathematics (OCIAM), Mathematical Institute, University of Oxford, Oxford, UK

[c] Oxford Centre for Integrative Systems Biology, Department of Physics, University of Oxford, Oxford, UK

[d] Smith School of Enterprise and the Environment, University of Oxford, UK.

[e] Speech, Language, and Hearing Science, University of Colorado, Boulder, Colorado, USA

[f] National Center for Voice and Speech, Denver, Colorado, USA

Emails: (A. Tsanas) tsanas@maths.ox.ac.uk, tsanasthanasis@gmail.com

(M. Little) littlem@maths.ox.ac.uk

(P. McSharry) patrick@mcsharry.net

(L. Ramig) Lorraine.Ramig@colorado.edu

Key words: Nonlinear speech signal processing, nonlinear regression and classification, Parkinson's disease, telemedicine, Unified Parkinson's Disease Rating Scale (UPDRS)

Short title for page headings: Statistical mapping of speech to a clinical Parkinson's disease metric

We have no conflict of interest. A. Tsanas is funded, in part, by Intel Corporation and by the Engineering and Physical Sciences Research Council (EPSRC).

Summary

The standard reference clinical score quantifying average Parkinson's disease (PD) symptom severity is the Unified Parkinson's Disease Rating Scale (UPDRS). At present, UPDRS is determined by the subjective clinical evaluation of the patient's ability to adequately cope with a range of tasks. In this study, we extend recent findings that UPDRS can be objectively assessed to clinically useful accuracy using simple, self-administered speech tests, without requiring the patient's physical presence in the clinic. We apply a wide range of known speech signal processing algorithms to a large database (~6,000 recordings from 42 PD patients, recruited to a six-month, multi-centre trial) and propose a number of novel, nonlinear signal processing algorithms which reveal pathological characteristics in PD more accurately than existing approaches. Robust feature selection algorithms select the optimal subset of these algorithms, which is fed into non-parametric regression and classification algorithms, mapping the signal processing algorithm outputs to UPDRS. We demonstrate rapid, accurate replication of the UPDRS assessment with clinically useful accuracy (about 2 UPDRS points difference from the clinicians' estimates, $p < 0.001$). This study supports the viability of frequent, remote, cost-effective, objective, accurate UPDRS telemonitoring based on self-administered speech tests. This technology could facilitate large-scale clinical trials into novel PD treatments.

1. Introduction

Parkinson's disease (PD) is a common neurodegenerative disorder with prevalence rates exceeding 100/100,000 (von Campenhausen *et al*. 2005). Furthermore, it is possible that these statistics underestimate the problem, since an additional 20% of people with Parkinson's (PWP) are not diagnosed (Schrag *et al*. 2002). Given that age is the single most important risk factor for PD onset, particularly after age 50 (Elbaz *et al*. 2002), and the fact that the population is growing older, these figures could rise further in the near future.

PD is believed to be due to substantial dopaminergic neuron reduction in a brain region known as the basal ganglia, and its aetiology is unknown (hence it is often referred to as *idiopathic* PD). *Parkinsonism* exhibits similar PD-like symptoms, but these can be attributed to known causes, such as drugs or exposure to neurotoxins. The constellation of PD symptoms includes tremor, rigidity and general movement disorders, as well as cognitive impairment (Pahwa and Lyons 2007). Speech disorders are amongst the earliest indicators of PD onset (Harel *et al*. 2004), and are reported in about 90% of PWP (Ho *et al.* 1998); moreover 29% of the patients themselves regard speech impairment as one of their most troublesome symptoms (Hartelius and Svensson 1994). In addition, there is ample empirical evidence for speech degradation as the disease progresses (Harel *et al*. 2004; Holmes *et al.* 2000; Skodda *et al.* 2009), typically attributed to reduced voice amplitude (*hypophonia*), and increased breathiness (noise) in the PWP's voice (Ho *et al*. 1998; Pahwa and Lyons 2007).

At present, there is no cure for PD, although medication and surgical intervention may alleviate some of the symptoms and improve quality of life for most (Singh *et al.* 2007). However, early diagnosis and frequent disease tracking are critical to maximizing the effect of treatment (Tolosa *et al.* 2009; Pahwa and Lyons 2007). PD symptom tracking is currently achieved via regular physical visits by the PWP to the clinic, and the *subjective* assessment of the subject's ability to perform a range of empirical tests as observed by expert clinical raters. Nevertheless, despite the clinicians' experience and the available guidelines, PD symptom assessment often varies between experts (*inter-rater variability*) (Ramaker *et al.* 2002; Post *et al.* 2005) accentuating the need for an *objective* clinical tool to track average PD symptom progression.

As part of the clinical assessment, the PWP's ability to complete the requested empirical tasks is mapped to a rating scale specifically designed to follow disease progression. Of the various rating scales for monitoring PD progression, the Unified Parkinson's Disease Rating

Scale (UPDRS) is the most widely used for quantifying symptom severity (Ramaker *et al.* 2002). For untreated patients the UPDRS comprises a total of 44 *sections* where each section spans the numerical range 0-4 (0 denotes healthy and 4 denotes severe symptoms), and the final UPDRS is the summation of all sections (numerical range 0-176, with 0 representing perfectly healthy individual and 176 total disability). The UPDRS consists of three *components*: (1) Mentation, behavior and mood (4 sections); (2) Activities of daily living (13 sections), assessing whether PWP can complete daily tasks unassisted; and (3) Motor (27 sections), addressing muscular control. We refer to all three components collectively as *total UPDRS*. The third component commonly referred to as *motor UPDRS*, includes the sections 18-44 and ranges from 0-108, with 0 indicating no motor symptoms (such as tremor, rigidity, posture, stability, bradykinesia) and 108 denoting total lack of motor control. *Speech* appears explicitly in two sections: once in section 5 (understandable speech – part of the second UPDRS component) and once in section 18 (expressive speech – part of the third UPDRS component), and ranges between 0-8 with 8 being unintelligible. The medical rater assesses the subject's speech performance (quantifying how understandable and expressive speech is) during casual discussion. Figure 1 presents succinctly the details of the UPDRS metric.

Telemonitoring-based health care is an emerging field combining medical care and Internet-enabled technology. On the one hand, it facilitates fast, frequent, remote tracking of disease progression, minimizing the need for regular and inconvenient visits to the clinic. On the other hand, it significantly alleviates the burden on national health systems of excessive workload and the large, associated costs of clinical human expertise. Recently, Intel Corporation's novel telemonitoring system, known as the *At-Home Testing Device* (AHTD), was developed (Goetz *et al.* 2009). This device facilitates remote, non-invasive self-administered tests, which are specifically designed to track PD progression and include manual dexterity and speech tests. The speech tests consist of *running speech* and *sustained vowel phonations*; in this study we concentrate on the latter. The use of sustained vowels, where the subject is requested to hold the frequency of phonation steady for as long as possible, builds on empirical evidence that healthy subjects can elicit steady phonation, whereas subjects with some form of vocal impairment cannot (Titze 2000). The use of sustained vowels to assess the extent of vocal symptoms avoids some of the known confounding effects of articulatory movement in running speech (Schoentgen and De Gucteneere 1995), and is therefore common in general speech clinical practice (Titze 2000).

Previous studies used speech signals aiming to separate PWP from healthy controls (Harel *et al.* 2004; Little *et al.* 2009), and in the past year some authors highlighted the importance

of exploring the topic of mapping speech signals to UPDRS (Skodda *et al.* 2009; Goetz *et al.* 2009) in future studies. Motivated by these studies, we have recently used a number of well known speech signal processing algorithms which are traditionally used by clinical speech scientists to characterize *dysphonias* (malfunctions in voice production) and demonstrated the feasibility of using statistical machine learning techniques to map the results of these algorithms (*features*) to motor-UPDRS and total-UPDRS (Tsanas *et al.* 2010a; Tsanas *et al.* 2010b).

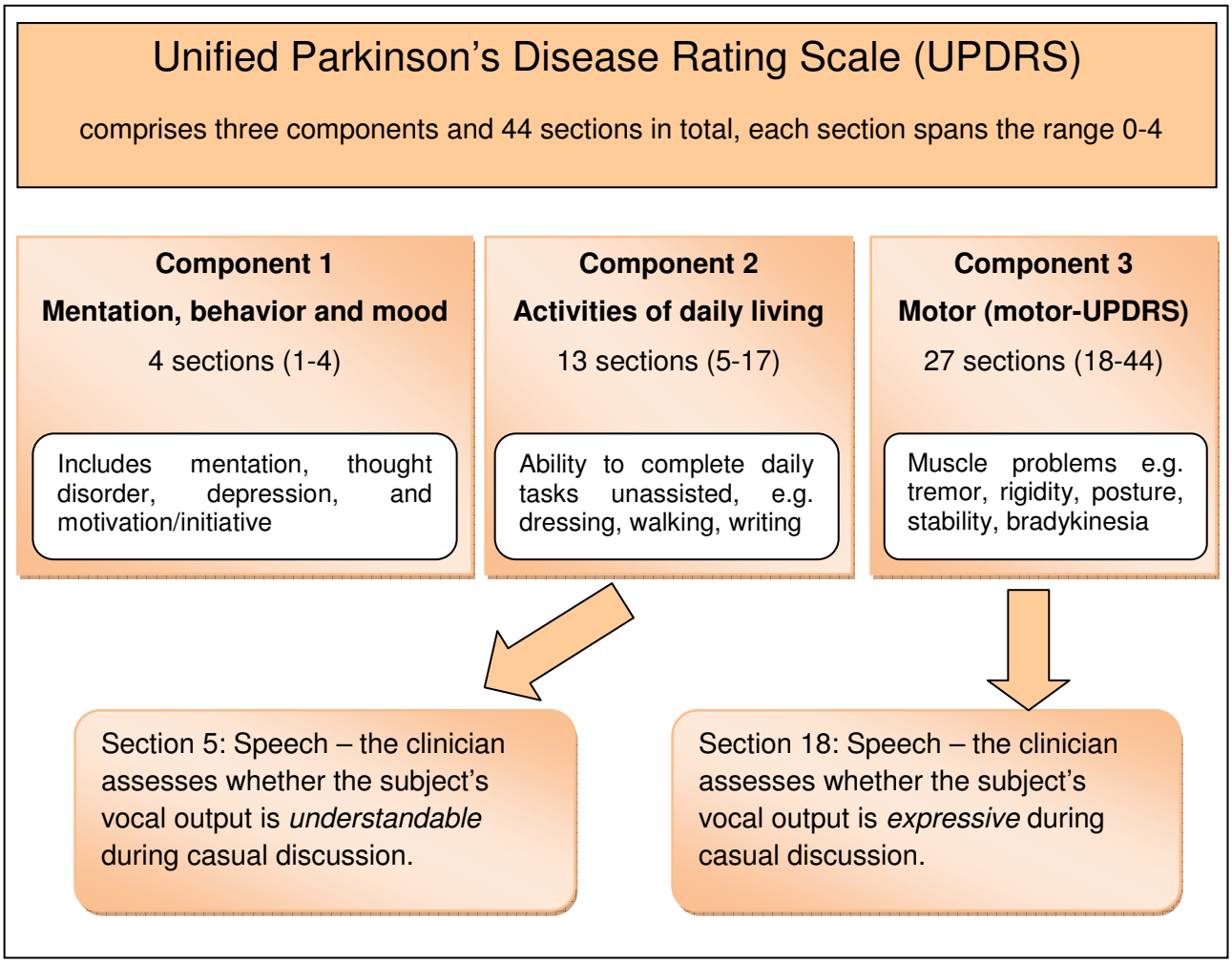

Figure 1. Overview of the clinical metric that quantifies average Parkinson's disease symptom severity, the Unified Parkinson's Disease Rating Scale (UPDRS). Speech appears explicitly twice.

In this study, we expand our analysis to introduce and investigate a range of speech signal processing algorithms which have not previously been used to characterize PD voices. Moreover, we present some novel nonlinear speech signal processing measures, which uncover many useful properties and characteristic patterns of PD dysphonia, that to-date,

remained concealed due to limitations of existing speech signal processing algorithms. In addition, we show that splitting the data into male and female data subsets (*data partitioning*) reveals distinct speech PD progression characteristics in males and females and this tentatively suggests different pathological patterns in these two groups. We demonstrate that we can replicate the clinicians' UPDRS estimates to within 2 points, that is, with *greater accuracy than the inter-rater variability* (4-5 UPDRS points) (Post *et al*. 2005). These new findings significantly improve on previous studies which introduced the concept of using speech signals to replicate the clinicians' UPDRS assessment, where the reported UPDRS accuracy was within 7.5 points.

This proposed *objective* machine learning framework using speech signals offers a promising approach to automating *subjective* UPDRS tracking, which would otherwise require the dedicated time of a clinical rater. This innovative approach is less cumbersome for patients since it reduces the need for frequent physical visits to the clinic. It is therefore also cost-effective for national health systems, and replicates the clinicians' estimates very accurately. We envisage this method being used to regularly and remotely track PD symptom progression by UPDRS, and facilitating large scale clinical trials into novel PD treatments. Lastly, the proposed signal processing features could be useful in affiliated research fields that use acoustic analysis of speech signals to assess various voice production pathologies.

2. Data

We use data collected in the study of Goetz *et al*. (2009), recently summarized in Tsanas *et al.* (2010a). In short, 52 subjects diagnosed with idiopathic PD within the previous five years at the time of a baseline clinical visit, were recruited into a trial of the AHTD. All subjects gave written informed consent, remained un-medicated for the six-month duration of the study and were asked to complete a range of tests weekly. Subjects were diagnosed with PD if they had at least two of the following symptoms: rest tremor, bradykinesia (slow movement), or rigidity, without evidence of other forms of Parkinsonism. No exclusion criteria related to specific PD symptoms (e.g. depression) were used. We disregarded data from 10 recruits – two that dropped out the study early, and a further eight that did not complete at least 20 valid study sessions during the trial period. Thus, this study concentrates on 42 PWP, and their details are summarized in Table 1.

Table 1: Summary of the AHTD data for the recruited male and female subjects.

	MALES (28 subjects)	FEMALES (14 subjects)
Age (years)	Mean ± standard deviation: 64.8 ± 8.1, min. 49, max. 78, median 65	Mean ± standard deviation: 63.6 ± 11.6, min. 36, max. 85, median 64
Weeks since PD diagnosis	Mean ± standard deviation: 63.0 ± 61.9, min. 1, max. 260, median 48	Mean ± standard deviation: 89.7 ± 81.2, min. 4, max. 252, median 60
Motor-UPDRS (baseline, 3-months, 6-months)	Mean ± standard deviation: (20.3 ± 8.5, 21.9 ± 8.7, 22.0 ± 9.2), min. (6, 6, 5), max. (36, 38, 41), median (21, 22, 20)	Mean ± standard deviation: (17.6 ± 7.4, 21.2 ± 10.5, 20.1 ± 9.4), min. (6, 6, 8), max. (32, 38, 38), median (18, 18.5, 19.5)
Total-UPDRS (baseline, 3-months, 6-months)	Mean ± standard deviation: (27.5 ± 11.6, 30.4 ± 11.8, 31.0 ± 12.4), min. (8, 7, 7), max. (54, 55, 54), median (27, 28.5, 26.5)	Mean ± standard deviation: (24.2 ± 9.1, 27.4 ± 12.1, 26.8 ± 10.8), min. (10, 7, 10), max. (42, 46, 49), median (25, 28, 24.5)

Table 2: Specifications of the At-Home Testing Device (AHTD) speech data collection interface.

AHTD instructions	Audible prompts prior to each test; visual prompts on the liquid crystal display, additional detailed help (text) available if needed
Microphone	High quality head-mounted, placed 5 cm from the subject's lips, Polarity: cardioid, typical dynamic range: 96 dB, 1 kHz at maximum sound pressure level, signal to noise ratio 58 dB, 1 kHz at 1 Pa, Frequency response: 100-13,000 Hz, Low frequency roll-off: 80 Hz, 18 dB/octave
Analogue-to-digital conversion	24 kHz at 16 bits resolution
Storage	Data recorded directly onto the AHTD USB data stick
Recording conditions	Subjects are required to be in a quiet place at home
Transmission	Data encrypted, transmitted over the internet to dedicated server

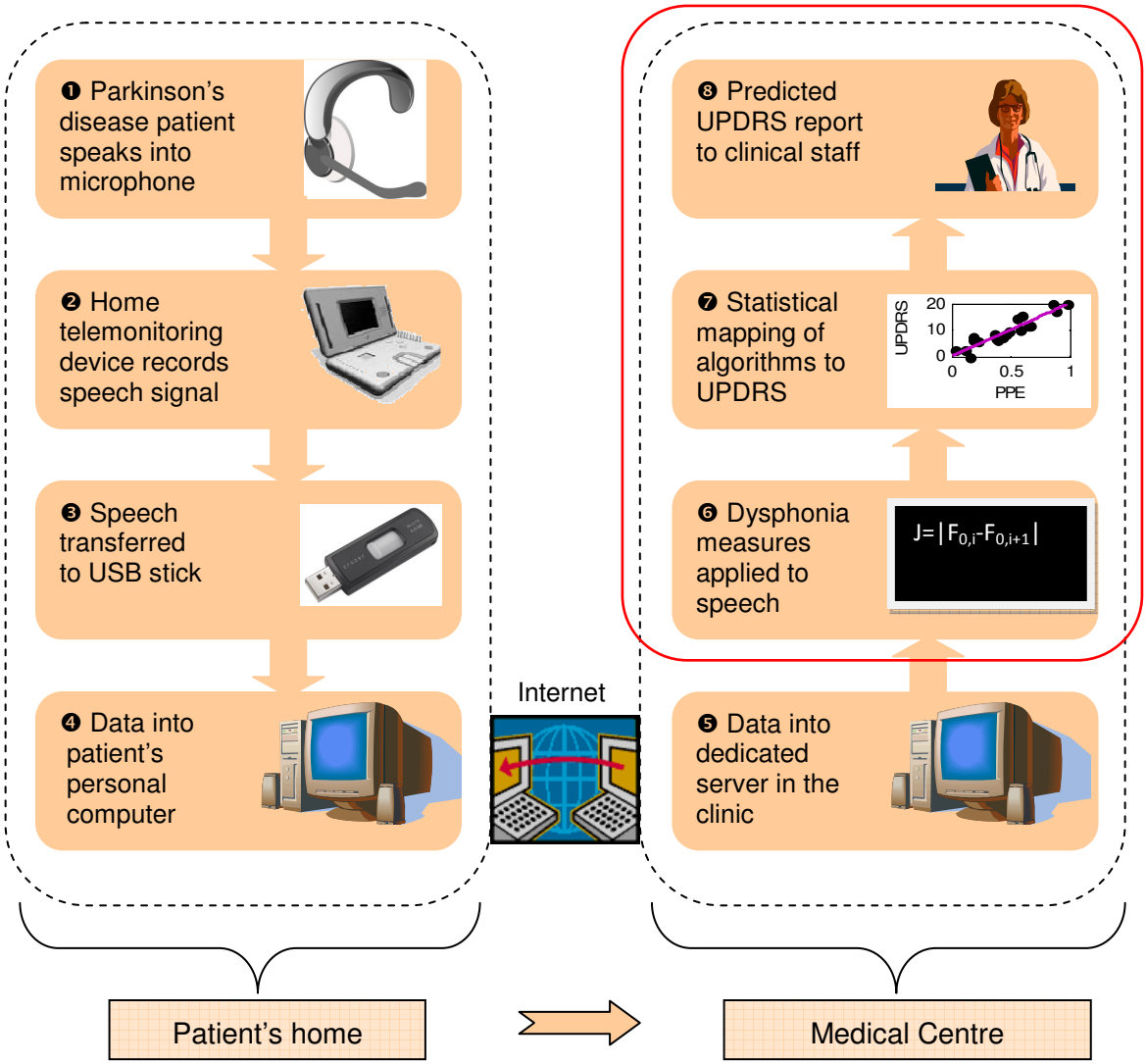

Figure 2 Schematic diagram of the steps from the data acquisition up to UPDRS estimation. The device that collects the data from the Parkinson's disease (PD) patient is known as the At-Home-Testing-Device (AHTD). The red box (steps 6-8) is the focus of this study.

A schematic diagram of the speech data acquisition process using the AHTD and the UPDRS estimation is presented in figure 2, and specifications of equipment are summarized in Table 2. The subjects in the study successfully completed a period of training in usage of the AHTD and used the device at their homes to self-collect the data. On each day the test was performed, the AHTD recorded six phonations: four at comfortable pitch and loudness and two at twice the initial loudness (but without shouting). The AHTD uses audible and visual prompts instructing the user to undertake specific tasks, including how to wear the head-mounted headset and the use of twice the initial loudness in the two final phonations. Although this latter aspect was not explicitly quantified, it has been empirically found that

paying conscious attention to speech articulation results in vocal performance improvement (Ho *et al*. 1999). Further details of the AHTD trial can be found in Goetz *et al.* (2009).

After initial screening to remove flawed phonations (too short, patient coughing, failure to capture phonation onset), we processed 5,875 sustained vowel "ahh…" signals. All signal processing and machine learning algorithms were implemented in the Matlab software package.

3. Methods

The methodology of this study can be succinctly described in three steps: 1) extracting features characterizing the underlying patterns of the speech signals using signal processing algorithms (*feature extraction*), 2) selecting a parsimonious subset of these features comprising relevant and minimally overlapping information with regard to UPDRS prediction (*feature selection*), and 3) mapping the feature subset to UPDRS using classification and regression methods (*statistical mapping*) in a standard *supervised learning* setup. Ultimately, we want to use the speech signals to replicate the clinicians' UPDRS assessment. In doing this, we tacitly assume that voice degradation is attributed solely to PD. It is conceivable that vocal performance could have been affected by confounding factors (for example emotional state) or pathological conditions (for example a disorder of voice production not related to PD). However, it is highly unlikely that these confounding factors affect more than a small minority of the AHTD subjects, thus contaminating only a few of the available recordings. Another source of error might be equipment tolerance. However, the speech data acquisition equipment is more than sufficient for the requirements of reliable speech signal processing (for details of the minimum requirements see (Titze 2000)), and thorough tests before the AHTD trial data acquisition process verified that the high-quality equipment used in the device lead to accurate recordings.

3.1 Feature extraction

The duration between two successive openings (or closures) of the vocal folds defines a *vocal fold cycle* (or simply *cycle*), where the *vocal fold oscillation pattern* (vocal fold opening and closure) is typically considered *nearly periodic* in healthy voices. That is, the intervals of time where the vocal folds are apart or in collision remain almost equal between

successive cycles. Speech scientists typically refer to those oscillation intervals as *pitch period* or *fundamental frequency* F_0 (reciprocal of pitch period – see figure 3). Whereas in healthy voices the vocal folds collide and remain together for a fixed portion of the cycle, in voice pathologies this pattern may be severely affected. In addition, a common manifestation of vocal impairment is incomplete *vocal fold closure*, resulting in excessive breathiness (noise). This imbalanced vocal fold movement also results in turbulent noise and the appearance of vortices in the airflow from the lungs, increasing the energy at higher energy components (Godino-Llorente *et al.* 2006). In general, people with voice disorders cannot elicit *steady* phonations (Titze 2000), and speech signal processing algorithms attempt to quantify this inefficiency at converting steady airflow from the lungs into stable voice.

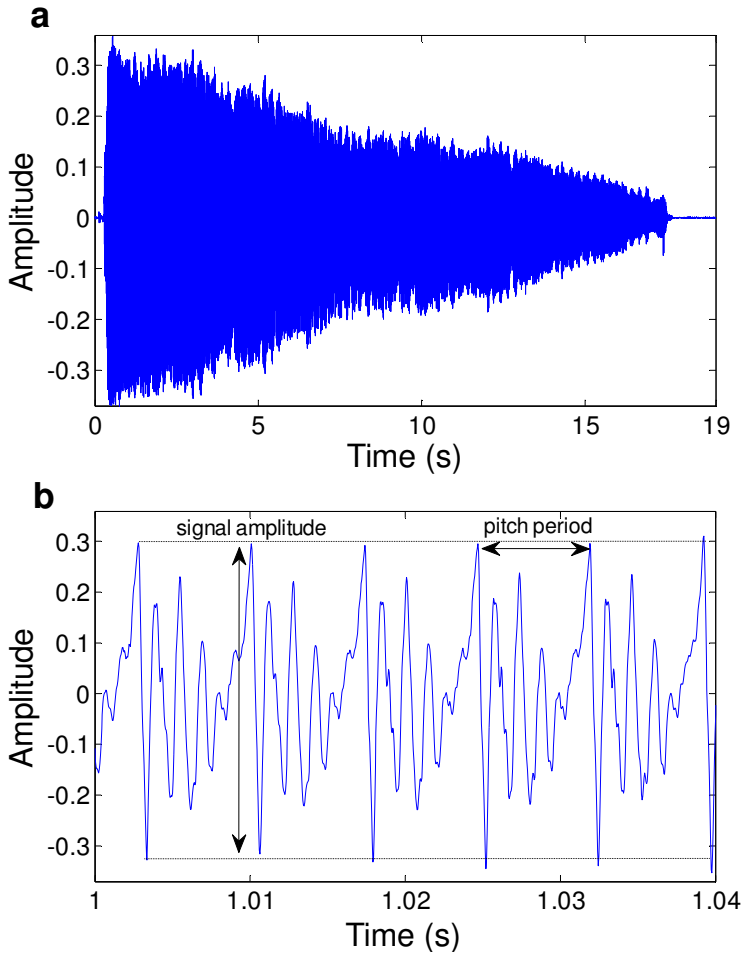

Fig. 3 (a) Typical sustained vowel phonation signal. (b) The same signal magnified in the time axis. The horizontal axes are time in seconds and the vertical axes amplitude (no units). Clear overall amplitude decay over the duration of the phonation can be seen in panel (a). A careful look at the magnified signal (b) reveals that it is not exactly periodic, a characteristic that many dysphonia measures aim to address.

The aim is to analyse the digitized acoustic signal using signal processing algorithms that take into account the pathophysiological implications outlined above, so that useful clinical information can be extracted. These algorithms are collectively known as *dysphonia measures* in the speech literature. Each of those measures is applied to each of the 5,875 recordings used in the study, resulting in a scalar value or a vector with a few entries per recording. Many algorithms work on *time windows* (small portions of the original speech signal). The output of those algorithms is then typically the average or some form of normalized average of the computed values on each of the time windows.

Previously, we had used the freely available Praat software package (Boersma and Weenink 2009) to extract 13 commonly-used measures (Tsanas *et al.* 2010a; Tsanas *et al.* 2010b) and three new measures we had proposed recently (Little *et al.* 2007; Little *et al.* 2009). In this study, all algorithms were implemented in Matlab using the equations described in the Electronic Supplementary Material (Section 1). In addition to the classical dysphonia measures, we introduce a range of novel nonlinear measures which we demonstrate convey important additional information useful in replicating the clinicians' UPDRS estimates. The outputs of the signal processing algorithms are concatenated into a feature vector which characterizes each of the 5,875 phonations.

3.2 Data exploration and statistical analysis

The UPDRS values of this study were obtained at baseline, three-month and six-month times in the trial, but the voice recordings were obtained weekly; therefore we need to obtain weekly UPDRS values to associate with each phonation. There is strong empirical evidence that *average* PD symptom progression in the *early stages* of the disease (up to about five years) is almost linear in *non-medicated* patients as observed in clinical metrics (Schüpbach *et al.* 2010; Maetzler *et al.* 2009). Therefore, given that the AHTD study recruits were in the early PD stages and remained non-medicated, a straightforward piecewise linear interpolation going exactly through the measured baseline, three-month and six-month motor-UPDRS and total-UPDRS scores is the most parsimonious and sensible approach to derive weekly values (Tsanas *et al.* 2010a; Tsanas *et al.* 2010b). The tacit assumption is that symptom severity did not fluctuate wildly within the three-month intervals in between which the UPDRS scores were obtained.

Correlation coefficients are the first quantities we explored in attempting to assess the

strength of association of the dysphonia measures with the linearly interpolated UPDRS values. The data was non-normal, so we used the non-parametric Spearman correlation coefficient. We also computed p-values (at the 95% level) of the null hypothesis against each dysphonia measure being uncorrelated with motor-UPDRS and total-UPDRS. In addition, we calculated the Spearman correlation coefficients *between* different dysphonia measures to assess the extent to which they contain overlapping information. We have also used the *mutual information* (MI) $I(X,Y)$, where X,Y are random variables (Cover and Thomas 2006), as a more inclusive, robust estimator of the association strength between the measures and UPDRS. The mutual information is non-negative, and is not upper bounded; therefore for ease of comparison we normalized $I(X,Y)$ by dividing it through with $I(Y,Y)$: hence, the reported mutual information in this study lies in the range zero (no dependence between X,Y) to one (X determines Y completely). Both the correlation coefficients and the mutual information are used to express the association strength (*relevance*) of each measure with UPDRS.

3.3 Feature selection

A ubiquitous problem in data analysis is the *curse of dimensionality*: the presence of a large number of features occludes the elucidation of useful patterns underlying the data, and is often detrimental in the subsequent learning process (see Section 3.4). This occurs because the required samples to adequately populate the feature space grow exponentially with the number of features, and typically is considerably more than the available data. Following the general principle of *parsimony*, which simply means that given several models with equal predictive power, we should prefer the model that uses the least number of features, it is desirable to reduce the number of features (hence produce a *sparse* model) in the analysis and still obtain an accurate estimate of the UPDRS. Selecting a subset of features may or may not improve the model's prediction accuracy; however it always enhances the model's *interpretability*. This is because we can infer the predominant characteristics of the dataset from the properties (*latent factors*) that the selected features represent, and a small number of features promote understanding of the causal relationship between those properties and UPDRS.

Searching through all possible combinations of features is unfeasible because it is computationally intractable in principle, giving rise to the need for computationally efficient

feature selection algorithms. We have used two generic, powerful feature selection methods: the *least absolute shrinkage and selection operator* (LASSO) (Tibshirani 1996), and a popular LASSO extension, the *elastic net* (Zou and Hastie 2005). Details of these algorithms and their promising *sparsity-promoting* properties can be found in (Tibshirani 1996; Zou and Hastie 2005; Hastie *et al.* 2009). For both algorithms we computed the entire regularization solution paths (Hastie *et al.* 2009).

3.4 Regression and classification: mapping dysphonia measures to UPDRS

The analysis in Section 3.2 provides preliminary indication of the association strength of each measure with UPDRS. However, the ultimate aim of this study is to combine the dysphonia measures to predict motor-UPDRS and total-UPDRS so that the absolute difference between the estimated and the linearly interpolated UPDRS is minimized. That is, we need to form a functional relationship $f(\boldsymbol{x}) = y$ which maps the dysphonia measures $\boldsymbol{x} = (x_1 \dots x_M)$, where M is the number of input variables, to the UPDRS output y. This is the classical *supervised learning* setup, which for the problem in question can be tackled using either *regression* or *classification* mapping techniques. Following the linear interpolation described earlier, the UPDRS spans the range of positive real values, i.e. $y \in \mathbb{R}^+$, which is what we use as the mapped quantity (also known as *outcome measurement* or *response variable*) in the regression scheme. For the classification schemes we used the rounded y scores and treat each integer UPDRS value as a different class.

Previous studies have shown the limitations of classical linear regression methods in this application (Tsanas *et al.* 2010a; Tsanas *et al.* 2010b), indicating that nonlinear methods may be more appropriate. In particular, we have experimented with *Classification and Regression Trees* (CART), and *Random Forests* (RF). Both CART and RF were tested working in both regression and classification modes.

CART was the method of choice in Tsanas *et al.* (2010a) because it has been described as the best off-the-shelf mapping algorithm in supervised learning contexts (Hastie *et al.* 2009). It partitions the feature space into hyper-rectangles, assigning a value to each of the hyper-rectangles that is as close as possible in value to the response variable in that region of the feature space (typically the mean or the median of the response values in that hyper-rectangle). This can be viewed as a tree growing process, where each partition splits in two branches. To avoid *overfitting*, i.e. capturing noisy fluctuations in the data at the expense of

the underlying structure of the mapping, an internal *pruning level* parameter is used to remove excessive detail in the partitioning of the feature space. The optimal pruning level value is typically determined by cross-validation. For further details on the advantages of the method and its mathematical foundations, we refer to Hastie *et al.* (2009).

A natural extension of CART is *random forests* (RF), a method comprising of many *de-correlated trees*, and can be thought of as *ensemble learning*, that is, integrating the 'opinion' of many *weaker* individual learners (Breiman 2001). The procedure is essentially the same as CART regarding the training of the trees (hyper-rectangle feature space partition described above); the only difference is that a random subset of the input features is chosen for each tree. The tree-growing process is the same as in CART, and there is no pruning; the prediction result of the RF learner is an average of the prediction from each tree. Breiman convincingly demonstrated that random forests are effective in various prediction tasks, whilst they do not overfit as more trees are added to the RF (Breiman 2001). For more information on RF we refer the reader to Hastie *et al.* (2009).

It is possible that partitioning the data may provide improved classification and regression accuracy in statistical machine learning applications. We partitioned the PWP according to gender, to investigate whether PD progression can be captured more accurately. That is, instead of building a 5,875×M matrix of feature vectors with all the data (*design matrix*), we used a design matrix of size 4,010×M for male and 1,865×M for female PWP. These design matrices contained no invalid or missing entries. Prior to feature selection, we have 132 dysphonia measures (i.e. initially, $M = 132$).

3.5 Cross validation and model generalization

We used 10-fold cross-validation to test the *generalization performance* of the learners used in this study. This represents our best estimate of UPDRS estimation performance on what we might expect on a new dataset, assuming the new dataset has similar characteristics to the AHTD data. Specifically, the initial dataset consisting of N (4,010 for males and 1,865 for females) phonations was split into a training subset of $0.9 \cdot N$ (3,609 and 1,679) phonations and a testing (*out of sample*) subset of $0.1 \cdot N$ (401 and 186) phonations. We repeated the process a total of 100 times, randomly permuting the data before splitting into training and testing subsets. Similar to our previous work (Tsanas *et al.* 2010a; Tsanas *et al.* 2010b), we

compared model performance on the basis of *mean absolute error* (MAE) for each of the 100 runs for the training and testing subsets:

$$MAE = \frac{1}{N}\sum_{i \in Q} |\hat{y}_i - y_i| \quad (1)$$

where $\hat{y}_i$ is the predicted UPDRS and y_i is the actual UPDRS for the i^{th} 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 cross-validation realisations were averaged.

4. Results

4.1 Data exploration

We began the exploration of the data by computing the relevance of speech features to UPDRS. Speech appears explicitly in two sections of the UPDRS, which can be combined to form the 'speech-UPDRS' quantity. Then, the relationships between speech-UPDRS and motor-UPDRS are ($p < 0.001$), Spearman $R = 0.464$, MI = 0.153 for males, and ($p < 0.05$), Spearman $R = 0.323$, MI = 0.199 for females. Similarly, the relationships between speech-UPDRS and total-UPDRS are ($p < 0.001$), Spearman $R = 0.552$, MI = 0.22 for males, and ($p < 0.05$), Spearman $R = 0.323$, MI = 0.168 for females. These preliminary statistical results offer good indication that speech and UPDRS are actually linked. Table 3 summarizes the dysphonia measures with the largest relevance to UPDRS for male PWP; similarly Table 4 for female PWP. All measures were significantly correlated ($p < 0.001$) with linearly interpolated motor-UPDRS and total-UPDRS, and some of these measures are quite strongly associated with UPDRS, particularly for the female PWP. In addition, figure 4 presents scatter plots of the most highly correlated dysphonia measures against UPDRS, giving a visual impression of the distribution of the dysphonia signal processing values and their relationship to UPDRS.

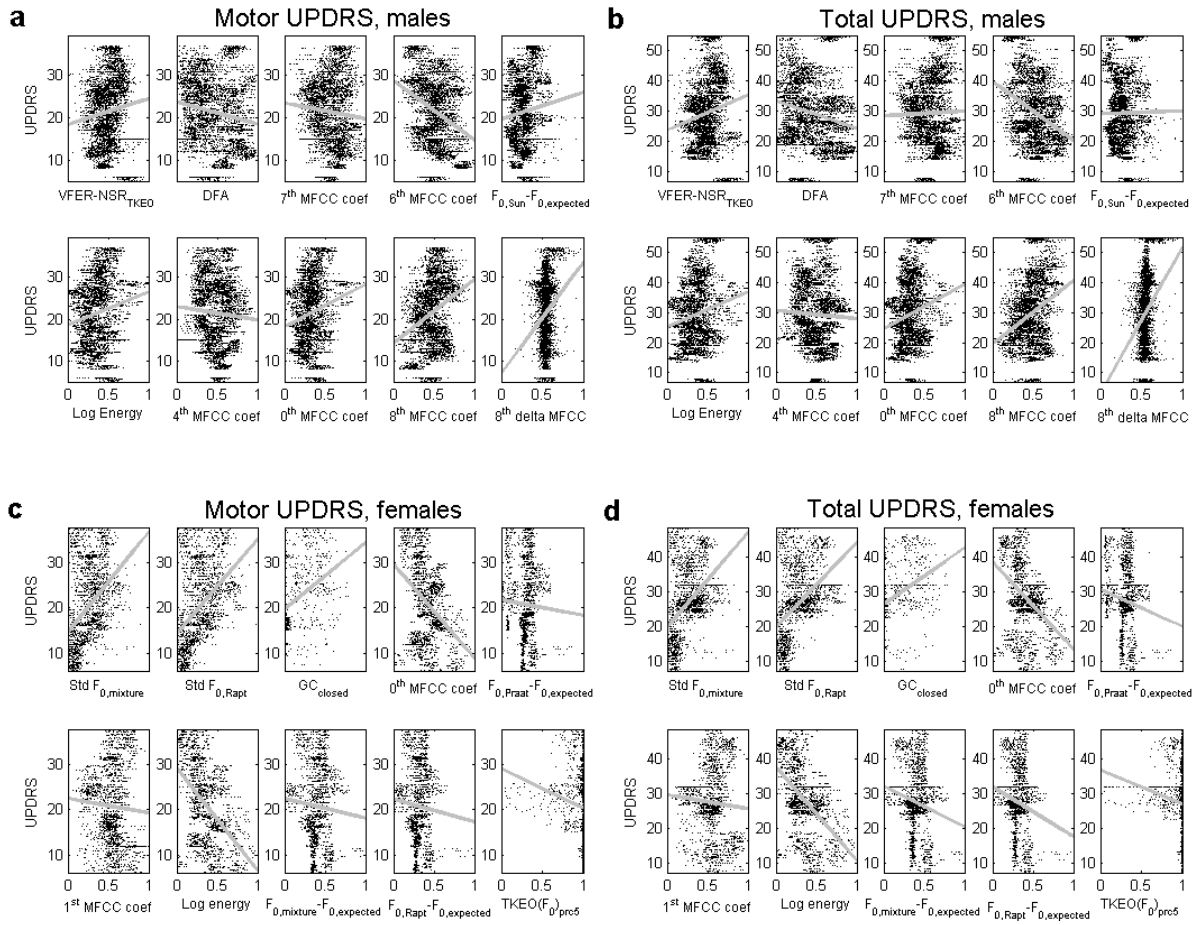

Fig. 4. Scatter plots of the most relevant dysphonia measures against motor UPDRS and total UPDRS for males and for females, using the measures presented in Tables 3 and 4. The horizontal axes are the normalized dysphonia measures and the vertical axes correspond to UPDRS. The gray lines are the best linear fit obtained using Iteratively Reweighted Least Squares – see (Tsanas *et al*. 2010) for details.

We can see that most of the times, large absolute correlation coefficient values correspond to large normalized MI values in Tables 3 and 4. However, some dysphonia measures have low absolute correlation coefficients and relatively large normalized MI (for example the 7th MFCC coefficient in Table 3). This indicates that those dysphonia measures are associated with UPDRS in a nonlinear *non-monotonic* way, which needs to be characterised using higher order moments (the Spearman correlation coefficient fails to quantify these relationships). Conversely, given two dysphonia measures (for example the VFER-NSR$_{TKEO}$ and the 8th delta MFCC coefficient in Table 3), a higher absolute value correlation coefficient might correspond to a lower normalized MI. This indicates that the extent of the association strength between the 8th delta MFCC coefficient and UPDRS can be adequately quantified using a monotonic relationship, whereas the extent of the association strength between the

VFER-NSR_{TKEO} and UPDRS relies more on higher order moments.

Table 3: Maximum relevance and correlations of dysphonia measures with UPDRS for males.

Measure	Description	Motor-UPDRS relevance and correlation		Total-UPDRS relevance and correlation	
		MI	Spearman R	MI	Spearman R
VFER-NSR_{TKEO}	Ratio of the sum of the log-transformed mean TKEO of the band-pass signals for frequencies >2.5 kHz to the sum of the mean TKEO of the band-pass signals for frequencies <2.5 kHz	0.105	0.159	**0.132**	0.187
DFA	Characterizes the extent of turbulent noise, quantifying its stochastic self-similarity (Little *et al.* 2007)	0.078	-0.162	**0.115**	-0.205
7^{th} MFCC coef	7^{th} Mel Frequency Cepstral Coefficient (Brookes 2006)	0.079	-0.066	**0.108**	0.0070
6^{th} MFCC coef	6^{th} Mel Frequency Cepstral Coefficient (Brookes 2006)	0.106	-0.277	**0.102**	-0.294
$F_{0,Sun} - F_{0,expected}$	Mean difference of the cycle-to-cycle F_0 estimate (extracted using Sun's algorithm) and the average expected F_0 in age- and sex-matched healthy controls	0.088	0.097	**0.101**	0.018
Log energy	Estimate of the logarithmic energy (Brookes 2006)	0.090	0.149	**0.099**	0.169
4^{th} MFCC coef	4^{th} Mel Frequency Cepstral Coefficient (Brookes 2006)	0.088	-0.082	**0.098**	-0.061
0^{th} MFCC coef	0^{th} Mel Frequency Cepstral Coefficient (Brookes 2006)	0.079	0.171	**0.099**	0.197
8^{th} MFCC coef	8^{th} Mel Frequency Cepstral Coefficient (Brookes 2006)	0.106	0.276	**0.095**	0.259
8^{th} delta MFCC coef	8^{th} delta Mel Frequency Cepstral Coefficient (First derivative of 8^{th} MFCC) (Brookes 2006)	0.073	0.181	**0.093**	0.205

The ranking was determined by the mutual information (MI) with the total UPDRS (for clarity, only the 10 most relevant measures are presented here). Relevance denotes the association strength of each feature with UPDRS expressed using the MI. The reported MI is normalized (i.e. MI lies between 0-1, where 0 denotes that UPDRS is independent on the dysphonia measure, and 1 indicates that UPDRS is completely determined by the dysphonia measure - see Section 3.2 for details). All results were rounded to the nearest third decimal digit. The UPDRS relevance and correlation columns are the MI where the probability density functions were computed with kernel density estimation with Gaussian kernels, and the Spearman non-parametric correlation coefficients between each measure and piecewise linearly interpolated motor and total UPDRS. All measures were statistically significantly correlated ($p < 0.001$) with motor-UPDRS and total-UPDRS. All speech signals from the male PWP were used to generate these results (N = 4,010 phonations). The F_0 subscript text refersto the algorithm used to extract it.

The overall impression we take from Tables 3 and 4 is that the most highly associated dysphonia measures with UPDRS are some of the MFCCs in males, and F_0-related measures for females. Specific MFCCs coefficients do not have particular *physical meaning*, but a more general interpretation is possible: lower MFCCs reflect the amplitude and envelope spectral fluctuations, and higher MFCCs convey mostly information about harmonic components (see the Electronic Supplementary Material for more information on MFCCs). The MFCCs in Table 3 are in the mid-range, and they are not easily interpretable since they

fall in neither category. We defer elaboration of the F_0-related measures for females for the Discussion.

Table 4: Maximum relevance and correlations of dysphonia measures with UPDRS for females.

Measure	Description	Motor-UPDRS relevance and correlation		Total-UPDRS relevance and correlation	
		MI	Spearman R	MI	Spearman R
Std $F_{0,mixture}$	Standard deviation of the extracted $F_{0,mixture}$	0.205	0.475	**0.216**	0.470
Std $F_{0,Rapt}$	Standard deviation of the extracted $F_{0,Rapt}$	0.174	0.437	**0.195**	0.434
GQ_{closed}	Standard deviation of the duration that the vocal folds remain closed	0.211	0.236	**0.195**	0.250
0th MFCC coef	0th delta Mel Frequency Cepstral Coefficient (Brookes 2006)	0.200	-0.327	**0.187**	-0.344
$F_{0,Praat} - F_{0,expected}$	Mean difference of the cycle-to-cycle F_0 estimate (extracted using Praat's algorithm) and the average expected F_0 in age- and sex-matched healthy controls	0.198	0.103	**0.176**	0.034
1st MFCC coef	1st delta Mel Frequency Cepstral Coefficient (Brookes 2006)	0.135	-0.047	**0.170**	-0.031
Log energy	Estimate of the logarithmic energy (Brookes 2006)	0.179	-0.458	**0.170**	-0.487
$F_{0,mixture} - F_{0,expected}$	Mean difference of the cycle-to-cycle F_0estimate (extracted using the mixture algorithm) and the average expected F_0 in age- and sex- matched healthy controls	0.181	0.019	**0.164**	-0.055
$F_{0,Rapt} - F_{0,expected}$	Mean difference of the cycle-to-cycle F_0 estimate (extracted using Rapt's algorithm) and the average expected F_0 in age- and sex-matched healthy controls	0.173	0.022	**0.158**	-0.054
$\Psi(F_0)_{prc5}$	5th percentile of the TKEO of the fundamental frequency values, obtained with the mixture algorithm	0.177	-0.411	**0.153**	-0.369

The ranking was determined by the mutual information (MI) with the total UPDRS (for clarity, only the 10 most relevant measures are presented here). Relevance denotes the association strength of each feature with UPDRS expressed using the MI. The reported MI is normalized (i.e. lies between 0-1, where 0 denotes that UPDRS is independent of the dysphonia measure, and 1 indicates that the UPDRS is completely determined by the measure - see Section 3.2 for details). All results were rounded to the nearest third decimal digit. The UPDRS relevance and correlation columns are the MI where the probability density functions were computed with kernel density estimation with Gaussian kernels, and the Spearman non-parametric correlation coefficients between each measure and piecewise linearly interpolated motor and total UPDRS. All measures were statistically significantly correlated ($p < 0.001$) with motor-UPDRS and total-UPDRS. All speech signals from the female PWP were used to generate these results (N = 1,875 phonations). The F_0 subscript text refers to the algorithm used to extract it.

4.2 Feature selection and statistical mapping of features to UPDRS

As described in Section 3.3, the LASSO and the elastic net can be used to determine the dysphonia measures that may be optimally included in a learner for UPDRS prediction. The feature selection process in this report used 10-fold cross validation (we experimented with 100 runs), where we recorded the selected features across all runs. The sparsity pattern of both the LASSO and the elastic net was very stable for the first 10 (and quite stable for the

first 15) selected features across the 100 realisations of the 10-fold cross validation. That is, the order of the initially selected features was almost the same across each cross-validation realisation used in feature selection. In Section 2.1 of the Electronic Supplementary Material we compare the 15 most important features selected by the two algorithms.

Table 5: Selected dysphonia measure subsets for males and females

MALES (33 dysphonia measures)					**FEMALES (33 dysphonia measures)**				
Dysphonia measure	Motor UPDRS		Total UPDRS		Dysphonia measure	Motor UPDRS		Total UPDRS	
	MI	R	MI	R		MI	R	MI	R
6^{th} MFCC coef	0.106	-0.277	0.102	-0.294	Log energy	0.179	-0.458	0.170	-0.487
8^{th} MFCC coef	0.106	0.276	0.095	0.259	Std $F_{0,Rapt}$	0.205	0.475	0.216	0.470
$VFER_{SNR,TKEO}$	0.077	-0.076	0.077	-0.108	10^{th} MFCC coef	0.112	0.239	0.107	0.250
$VFER_{mean}$	0.076	0.154	0.089	0.13	PPE	0.118	0.436	0.105	0.396
8^{th} delta MFCC	0.073	0.181	0.093	0.205	12^{th} MFCC coef	0.094	0.204	0.088	0.261
12^{th} delta MFCC	0.048	0.172	0.054	0.167	$IMF_{SNR,TKEO}$	0.075	-0.127	0.067	-0.067
0^{th} MFCC coef	0.079	0.171	0.097	0.197	8^{th} MFCC coef	0.114	-0.341	0.092	-0.255
2^{nd} MFCC coef	0.082	-0.149	0.084	-0.182	11^{th} MFCC coef	0.078	0.127	0.100	0.187
3^{rd} MFCC coef	0.071	0.091	0.077	0.067	$IMF_{NSR,SEO}$	0.099	-0.117	0.065	-0.058
2^{nd} delta MFCC	0.047	0.130	0.050	0.125	GNE_{mean}	0.090	0.035	0.086	-0.062
3^{rd} delta MFCC	0.046	0.169	0.054	0.161	3^{rd} delta MFCC	0.070	0.149	0.064	0.119
Std $F_{0,Sun}$	0.046	0.144	0.050	0.129	HNR_{std}	0.072	0.224	0.066	0.195
9^{th} MFCC coef	0.075	-0.194	0.073	-0.153	5^{th} MFCC coef	0.113	0.173	0.115	0.188
7^{th} MFCC coef	0.079	-0.066	0.108	0.007	2^{nd} delta MFCC	0.055	0.172	0.056	0.206
4^{th} delta MFCC	0.041	0.001	0.044	0.007	$GNE_{SNR,TKEO}$	0.036	0.038	0.042	0.033
$GNE_{SNR,TKEO}$	0.023	0.074	0.024	0.089	10^{th} delta MFCC	0.071	-0.064	0.066	-0.079
$Shimmer_{A0,abs}$	0.042	-0.079	0.058	-0.135	GQ_{open}	0.061	0.256	0.057	0.248
$\Psi(F_0)_{25th\ percentile}$	0.074	-0.136	0.078	-0.056	GQ_{closed}	0.211	0.236	0.194	0.25
$IMF_{SNR,TKEO}$	0.045	-0.122	0.054	-0.151	4^{th} MFCC coef	0.19	0.329	0.140	0.242
$Shimmer_{PQ1,K=5}$	0.041	-0.065	0.056	-0.113	$\Psi(F_0)_{95th\ percentile}$	0.162	0.413	0.137	0.361
$Shimmer_{PQ3,K=11}$	0.043	-0.071	0.057	-0.116	$OQ_{5\text{-}95\ percentile}$	0.005	-0.216	0.001	-0.231
11^{th} MFCC coef	0.081	-0.006	0.070	0.021	6^{th} delta MFCC	0.073	0.152	0.066	0.086
Jitter-$F_{0,abs}$	0.061	0.103	0.064	0.045	Std $F_{0,Praat}$	0.146	0.352	0.132	0.316
$Shimmer_{dB}$	0.040	-0.066	0.054	-0.113	DFA	0.115	-0.059	0.094	-0.023
$GNE_{NSR,TKEO}$	0.035	0.098	0.033	0.11	$VFER_{SNR,SEO}$	0.130	-0.253	0.084	-0.175
RPDE	0.040	0.003	0.044	0.064	Std $\Psi(F_0)$	0.170	0.325	0.152	0.269
5^{th} MFCC coef	0.082	0.010	0.081	-0.039	$VFER_{SNR,TKEO}$	0.085	-0.143	0.086	-0.112
HNR_{std}	0.068	0.058	0.086	0.134	5^{th} delta MFCC	0.052	0.075	0.059	0.073
$Jitter_{pitch\ period\%}$	0.048	0.070	0.052	0.039	7^{th} MFCC coef	0.086	0.036	0.077	0.044
13^{th} delta MFCC	0.038	0.114	0.043	0.134	9^{th} MFCC coef	0.084	0.157	0.073	0.147
DFA	0.078	-0.162	0.112	-0.205	3^{rd} MFCC coef	0.151	-0.132	0.117	-0.058
$VFER_{NSR,TKEO}$	0.105	0.159	0.132	0.187	6^{th} MFCC coef	0.169	0.137	0.145	0.084
12^{th} delta-delta MFCC	0.035	0.066	0.049	0.058	$\Psi(A_0)_{75th\ percentile}$	0.078	0.067	0.072	0.089

The order of the features in the subsets is the order with which they were selected in the LASSO algorithm (features that were initially selected and subsequently dropped in the LASSO path are not included). The selected feature subsets were determined using the one standard error rule (see text for details). The Table also presents the mutual information (MI) and Spearman *R* (relevance and correlation) of the selected features with respect to the motor-UPDRS and total-UPDRS. The reported MI is normalized (i.e. MI lies between 0-1, where 0 denotes that UPDRS is independent on the dysphonia measure, and 1 indicates that the UPDRS is completely determined by the dysphonia measure - see Section 3.2 for details). Descriptions of the dysphonia measures appear in Section 1 of the Electronic Supplementary Material.

Then, we used one feature subset at a time (experimenting with the feature subsets selected by the LASSO or the elastic net) as input to the CART and RF learners to train and test each of the four learners' performance. Additionally, all the dysphonia measures were used as inputs into the learners in order to have a (potentially over-complex) MAE benchmark against which we could compare our findings. The pruning level of the CART learners was determined by manual checks to minimize the MAE. By default, we used 500 trees in the RF learners.

In order to select the best feature subset, we have used the "one-standard-error" rule (Hastie *et al.* 2009): we pick the most parsimonious subset in which the MAE is no more than one standard deviation above the MAE of the best subset. The selected feature subsets for males and females are summarized in Table 5. In all cases, the RF working in classification mode outperformed the other learners. Table 6 presents the out-of-sample MAE using the RF learner in classification mode for the feature subsets of Table 5, and compares these findings with those in Tsanas *et al*. (2010a) and Tsanas *et al*. (2010b). The generalization ability of the models is verified by the fact that the in-sample and out-of-sample errors were similarly low.

Table 6: Summary of the Mean Absolute Error (MAE) results of this study, and comparison with the results of previous studies.

Measures	MAE for motor-UPDRS	MAE for total-UPDRS
Selected feature subset for *males* in Table 5	1.62 ± 0.17	1.96 ± 0.23
Selected feature subset for *females* in Table 5	1.72 ± 0.16	2.20 ± 0.21
Selected feature subset in Tsanas *et al*. (2010a)	5.95 ± 0.19	7.52 ± 0.25
Selected feature subset in Tsanas *et al*. (2010b)	6.57 ± 0.16	8.38 ± 0.23

The reported MAE results were obtained with the Random Forests (RF) working in classification mode. The errors are reported in the form mean ± standard deviation. In Tsanas *et al*. (2010a) and Tsanas *et al*. (2010b) we had pooled together all the available phonations (no separation between male and female groups). The inter-rater variability (difference in clinical symptom assessment between trained clinicians) is about 4-5 UPDRS points (Post *et al*. 2005) and the results in this study demonstrate, for the first time, that a machine learning approach can do better than this benchmark.

We use the Wilcoxon rank sum test to demonstrate the significance of these findings by comparing the UPDRS results obtained using the methodology of this study against some benchmarks. We compared the distribution of the MAE for motor-UPDRS and total-UPDRS

against the MAE that are obtained using the mean motor-UPDRS and mean total-UPDRS (which are used as benchmarks, respectively) for males and for females. The null hypothesis is that the medians of the distributions are equal. The Wilcoxon rank sum test rejected the null hypothesis and the results are statistically significant ($p < 0.001$) for all four cases. In addition, we use as another benchmark the UPDRS value for each subject at baseline (that is, the UPDRS estimate is assumed constant for each subject at the baseline score), and compute the MAE distributions of motor-UPDRS and total-UPDRS by using this value. In this case, the null hypothesis is that the medians of the MAE distributions using the methodology of this study, and the MAE distributions using the baseline value for the individuals are equal. The Wilcoxon rank sum test rejected the null hypothesis and the results are statistically significant ($p < 0.001$) for all four cases.

With the exception of Tsanas *et al*. (2010a) and Tsanas *et al*. (2010b), we are not aware of any previous studies that have focused on replicating the average PD symptom severity when this is quantified by a clinical metric, such as the UPDRS. A recent study has attempted to replicate three aspects of the UPDRS metric (tremor, bradykinesia, and dyskinesia), using accelerometers (Patel *et al*. 2009). We refer to the Electronic Supplementary Material for details and a comparison of the results using the methodology of this study and Patel *et al*. (2009) in replicating the clinical evaluation (UPDRS assessment by the clinical rater) of those three elements. Not surprisingly, it appears that accelerometers are better suited compared to speech signals to replicate the clinicians' assessment of average severity in those three motor symptoms. Although these three elements are important, they do not encompass the breadth of PD symptoms which are expressed in the diverse UPDRS metric, and therefore do not actually reflect the average PD symptom severity which we try to quantify in our work.

4.3 Six month UPDRS tracking for the AHTD trial

So far, we have focused on randomly selecting phonations and estimating the UPDRS without working on specific individuals for a period of time (UPDRS *prediction*). In this Section, we aim to test the model's ability for UPDRS *tracking* (weekly UPDRS estimation of an individual for the six month duration of the trial using the speech recordings). One approach is to train the learner using the dysphonia measures computed from all subjects without including the dysphonia measures from the specific subject whose UPDRS we want to predict. However, this is a very unstable scheme due to the finiteness of the data (there are

only 42 subjects in the AHTD trial), and we elaborate further on this issue in the Discussion. For that reason, we have used the UPDRS tracking approach that we describe next.

On every day the PWP took the AHTD tests, six sustained vowel phonations were recorded. Thus, as a proxy for leaving out all the dysphonia measures from a single subject for the 6-month duration of the AHTD trial (~140 speech signals × M dysphonia measures), we can leave out the dysphonia measures derived from one of the weekly tests, and test the learner's out-of sample tracking ability using these dysphonia measures (~25 × M). However, we have noticed that our algorithms occasionally deliver quite large UPDRS differences using the out-of sample dysphonia measures derived from each of the six sustained vowel tests of individuals which were captured on the same day. This suggests that spurious artefacts pertaining to one or more of the six weekly recorded phonations may not be representative of the weekly UPDRS estimate of the patients. Therefore, we propose training the learner using the dysphonia measures from all the sustained vowel phonations of all patients, with the exception of the dysphonia measures derived from the first of each of the weekly phonations for a selected individual (about 20-25), which are used for testing. Subsequently, we repeat the same methodology training the system with all the dysphonia measures from all patients, excluding the dysphonia measures derived from the selected individual involving successively either the second, third, fourth, fifth or sixth sustained vowel phonation test. The six weekly out-of-sample MAE results are then averaged, resulting in a single UPDRS estimate. Our experiments suggest that the scheme with weighting the average UPDRS estimates from the dysphonia measures of the six weekly phonations is a more robust method compared to randomly selecting the dysphonia measures computed from one of the six weekly phonations.

Figure 5 presents the UPDRS tracking of a male and a female PWP using the combination of the best feature subset and RF working in classification mode. We have purposefully chosen male and female PWP with uncharacteristic UPDRS patterns (whereas the norm for PWP is progressive increase in symptom severity) to demonstrate that the proposed methods can follow larger, unexpected UPDRS changes. The actual UPDRS of the presented male PWP increased slightly in the 3-month visit and subsequently reduced on the 6-month visit, whereas the female PWP shown here is the subject with the most irregular UPDRS pattern in the AHTD trial (sharp UPDRS increase in the 3-month visit and subsequent sharp decrease in the 6-month visit). The female subject in figure 5b is the individual we have used previously (Tsanas *et al*. 2010a). Inspection of figure 5c, 5d and the tracking figure of Tsanas *et al*. (2010a) verifies the superiority of the approach developed in the current study in remotely

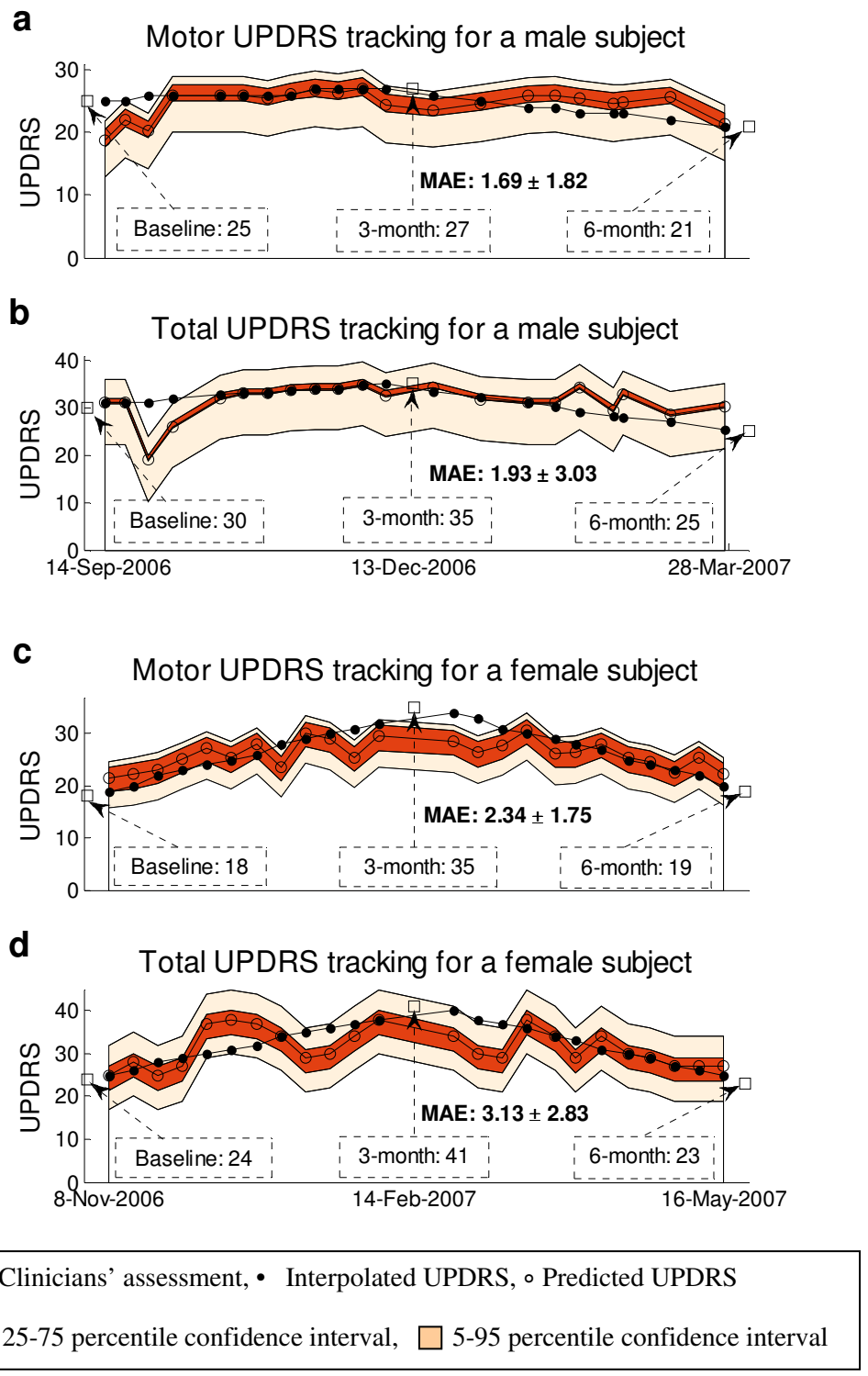

Figure 5 Motor-UPDRS and total-UPDRS tracking over the 6-month trial period for a male and a female subject with irregular UPDRS pattern. The 'baseline', '3-month' and '6-month' UPDRS scores are shown. The out-of-sample MAE and the standard deviation of MAE computed for the subjects presented in this figure are also quoted. The computation of the out-of-sample MAE and the confidence intervals reported in this figure were estimated from the average MAE of the six weekly error estimates throughout the six month duration of the trial for the specific individual.

following UPDRS symptom severity when this is expressed in UPDRS terms. We remark that the proposed models replicate quite accurately the linearly interpolated motor-UPDRS and total-UPDRS scores in figure 5. Generally, UPDRS increases monotonically for most of the patients, and the algorithm's tracking is even more precise in those cases.

5. Discussion

We have investigated the potential for using speech signals to estimate average PD progression with the standard reference clinical score, UPDRS. We stress that this study focused on PD *telemonitoring* and not PD *diagnosis*, which is a more difficult and subtle problem (to qualify as a diagnostic tool the methodology of this study should be applied in datasets that include healthy controls and, in addition, subjects with various neurological disorders that typically present PD-like symptoms). A wide range of known and novel speech signal processing algorithms (collectively known as dysphonia measures) have been implemented in order to uncover potentially concealed patterns in the PWP's voice and establish a functional mapping of these patterns to UPDRS. We have experimented with feature selection algorithms, aiming to select a parsimonious model with good prediction accuracy. The out of sample MAE were 1.6 points for males and 1.7 points for females for the motor UPDRS (which spans the range 0-108), and 2.0 points for males and 2.2 points for females for the total UPDRS (which spans the range 0-176), suggesting that the proposed methodology can accurately replicate the linearly interpolated UPDRS scores based on clinicians' subjective ratings. The new MAE results drastically improve upon Tsanas *et al.* (2010a) and Tsanas *et al.* (2010b) where the UPDRS was estimated to within 7.5 points. The improvement in the UPDRS estimation of this study is attributed to two factors: a) more sophisticated speech signal processing algorithms which uncover novel PD dysphonia patterns, b) the use of random forests, which clearly outperform CART in this application. We address each of these points later. We stress that we can replicate the clinicians' UPDRS estimates with accuracy that is considerably greater than the inter-rater variability (4-5 UPDRS points) (Post *et al*. 2005), a benchmark clinicians might want to refer to. These promising new results could convince more clinicians about the practical effectiveness of the proposed approach, and consequently lead to the adoption of the AHTD in larger clinical trials.

We started the exploration of the data by combining the two UPDRS sections with explicit "Speech" headings to form a composite speech-UPDRS score, and reported the association strength of speech-UPDRS with motor- and total-UPDRS. These results are built upon the idea that slight changes in the voice reflect some change in PD symptom severity. It is also highly likely that speech changes occur due to natural biological variation since humans do not produce identical outputs under identical conditions. Such sources of intrinsic variation in voice are, however, irrelevant to the systematic component of the relationship between voice and PD symptom severity: as we have demonstrated in this study and others, such intrinsic biological variability does not preclude prediction of PD symptom severity. It would however be of interest to understand such intrinsic biological variability of the voice for other purposes. The results of this study provide good statistical evidence that speech impairment and average, overall PD symptom severity are inherently linked, and intuitively justify the premise that UPDRS can be predicted by analysing speech signals alone.

Previous studies had only computed some of the commonly used dysphonia measures to investigate the potential of using sustained vowels to track average PD symptom progression. In this study, we have significantly reinforced earlier findings using additional speech signal processing algorithms, and proposing a number of novel algorithms which are able to detect previously hidden patterns in PWP's speech degradation. The new measures rely mainly in the physiological understanding that pathological voices exhibit increased tremor and high-frequency noise, and attempt to quantify these characteristics using energy and entropy concepts. The fact that the feature selection algorithms showed heavy bias towards selecting the non-classical measures is compelling evidence that these new measures quantify clinically useful information in PD voices which may not be captured by the classical dysphonia measures. We elaborate further on the issue of dysphonia measures in PD in the discussion Section of the Electronic Supplementary Material.

Interestingly, our experiments demonstrate that there are substantially different PD effects in the voices of male and female PWP. The mutual information and correlation coefficients for males in Table 3 and females in Table 4 reveal some interesting, and slightly surprising attributes. In particular, measures directly extracted from the fundamental frequency (both the standard deviation of the estimated F_0 and the absolute difference to the population average F_0 for matched healthy controls) appear strongly associated with UPDRS in females but apparently there is no similar distinctive pattern for males. We had previously reported that PPE, a measure which relies on the log-transform of the fundamental frequency, is one of the most important measures for predicting UPDRS (Tsanas *et al*. 2010). In fact, we have now established that this is

because PPE is an excellent predictor for UPDRS tracking in females, but is quite ineffective in males. Ultimately, the gender differentiation supports a tentative physiological conclusion: that the underlying processes of degradation in PD speech may be different in men and women. Moreover, the association strength of the dysphonia measures with UPDRS is much larger in females (Tables 3 and 4). In brief, we speculate this is because there is a distinct signature (pattern) characterising voice pathologies in females, whereas this pattern is masked in males due to the physiology of natural male voice production. Since higher fundamental frequencies tend to have lower perturbations (Baken and Orlikoff 2000), and given that women have higher average F_0 (Titze 2000), it is plausible that even slight distortions in vocal performance (for example aperiodic F_0) reflect voice pathology in females with high probability, whilst similar distortions in males' vocal performance can be attributed (at least partly) to normal vibrato. Thus, voice degradation quantified using some of the dysphonia measures (particularly those related to F_0) could represent general symptom degradation in females, whereas similar quantification of the voice perturbations in males could be part of the variability in normal voice production mechanisms.

We have experimented with nonlinear, nonparametric learners: CART and RF. We have used CART and RF working in both regression and classification modes, since the problem tackled in this study is amenable to both interpretations. In all simulations, RF outperformed CART, typically in excess of 1 UPDRS point. Our study agrees with Breiman's findings (Breiman 2001) that RF perform better in classification mode. The reported MAE estimates come from the 100 runs 10-fold cross-validation scheme and reflect our best estimate of the *asymptotic out-of-sample* prediction error given the available data. As we have argued previously (Tsanas *et al.* 2010a), the reliability of the cross-validation implicitly assumes independence between samples, which may be violated since we have typically about 140 samples from each of the 42 patients, and approximately 6,000 samples overall. However, any *patient-specific* validation scheme is unstable because there is not enough hold-out data to form reliable estimates of the learners' performance. This was verified in our experiments with a leave-one-patient-out cross validation scheme, where the standard deviations around the computed MAE were almost as large as the error. A simple test that goes some way towards determining whether the samples are truly independent is to use as an additional input feature (along with the selected subset of the dysphonia measures) the patient index: if there is large dependency between samples from the same patient, the out-of-sample MAE of the learners will be noticeably reduced. In doing this simple experiment we noticed a

marginal MAE reduction of about 0.2 UPDRS points, which is statistically insignificant. This evidence supports the interpretation that there is no strong dependence between samples from each patient.

Telemonitoring in healthcare is fast emerging, and is particularly important for PWP because it is often extremely awkward for those patients to make frequent visits to the clinic. Our findings could be useful in clinical trials, offering a novel approach to tracking average PD symptom severity by UPDRS remotely, and at frequent intervals. We envisage this technology finding application in future clinical trials of novel treatments which will require high-frequency, remote, and very large study populations.

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.

Declaration

We have no conflict of interest. A. Tsanas is funded, in part, by Intel Corporation, and in part by the Engineering and Physical Sciences Research Council (EPSRC).

References

Baken, R.J., and Orlikoff, R.F. 2000 Clinical measurement of speech and voice, 2nd edition, San Diego: Singular Thomson Learning

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

Breiman, L. 2001 Random Forests, *Machine Learning*, **45**, 5-32 (doi: 10.1023/A:1010933404324)

Brookes, M. 2006 VOICEBOX, Speech Processing Toolbox for Matlab, retrieved from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 2006

Chatfield, C. 2004 *The Analysis of Time Series: An Introduction*, 6th edition, London, Chapman & Hall/CRC

Cover T.M., and Thomas, J.A. 2006 *Elements of information theory*, 2nd edition, Wiley-interscience

Elbaz, A., Bower, J.H., Maraganore, D.M., McDonnell, S.K., Peterson, B.J., Ahlskog, J.E., Schaid D.J., and Rocca, W.A. 2002 Risk tables for parkinsonism and Parkinson's disease, *Journal of Clinical Epidemiology*, **55**, 25-31 (doi: 10.1016/S0895-4356(01)00425-5)

Godino-Llorente, J.I., Gomez-Vilda, P., Blanco-Velasco, M. 2006 Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters, *IEEE Transactions on Biomedical Engineering*, **53**,1943-1953

Goetz, C.G., Stebbins, G.T., Wolff, D., DeLeeuw, W., Bronte-Stewart, H., Elble, R., Hallett, M., Nutt, J., Ramig, L., Sanger, T., Wu, A.D., Kraus, P.H., Blasucci, L.M., Shamim, E.A., Sethi, K.D., Spielman, J., Kubota, K., Grove, A.S., Dishman, E., Taylor, C.B. 2009 Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device, *Movement Disorders*, **24**, 551-556 (doi: 10.1002/mds.22379)

Harel, B., Cannizzaro, M., and Snyder, P.J. 2004 Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study, *Brain and Cognition*, **56**, 24–29 (doi:10.1016/j.bandc.2004.05.002)

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

Hastie, T., Tibshirani, R., Friedman, J. 2009 *The elements of statistical learning: data mining, inference, and prediction*, 2nd edition, Springer, New York, USA

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

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

Holmes, R.J., Oates, J.M., Phyland, D.J., and Hughes, A.J. 2000 Voice characteristics in the progression of Parkinson's disease, *Int J Lang Comm Dis*, **35**, 407-418 (doi: 10.1080/136828200410654)

Little, M.A., McSharry, P.E., Roberts, S.J., Costello, D., Moroz, I.M. 2007 Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection, *Biomedical Engineering Online*, **6:23** (doi: 10.1186/1475-925X-6-23)

Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., Ramig, L.O. 2009 Suitability of dysphonia measurements for telemonitoring of Parkinson's disease, *IEEE Transactions Biomedical Engineering*, **56**, 1015-1022 (doi: TBME-00342-2008)

Maetzler, W., Liepelt, I., Berg, D. 2009 Progression of Parkinson's disease in the clinical phase: potential markers, *Lancet Neurology*, **8**, 1158-1171 (doi: 10.1016/S1474-4422(09)70291-1)

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

Patel, S., Lorincz, K., Hughes, R., Huggins, N., Growdon, J., Standaert, D., Akay, M., Dy, J., Welsh, M., and Bonato, P. 2009 "Monitoring Motor Fluctuations in Patients With Parkinson's Disease Using Wearable Sensors", *IEEE Transactions on Information technology in biomedicine*, Vol. **13** (6), pp. 864-873 (doi: 10.1109/TITB.2009.2033471)

Post, B., Merkus, M.P., de Bie, R.M.A., de Haan, R.J., and Speelman, J.D. 2005 Unified Parkinson's Disease Rating Scale Motor Examination: Are Ratings of Nurses, Residents in Neurology, and Movement Disorders Specialists Interchangeable?, *Movement Disorders*, **20**, 1577-1584 (doi: 10.1002/mds.20640)

Ramaker, C., Marinus, J., Stiggelbout A.M., and van Hilten, B.J. 2002 Systematic evaluation of rating scales for impairment and disability in Parkinson's disease, *Movement Disorders*, **17**, 867-876 (doi: 10.1002/mds.10248)

Schoentgen J. and De Gucteneere, R. 1995 Time series analysis of jitter, *Journal of Phonetics*, **23**, 189-201 (doi:10.1016/S0095-4470(95)80042-5)

Schrag, A., Ben-Schlomo, Y., Quinn, N. 2002 How valid is the clinical diagnosis of Parkinson's disease in the community?, *Journal of Neurology, Neurosurgery Psychiatry*, **73**, 529-535 (doi:10.1136/jnnp.73.5.529)

Schüpbach, M.W.M., Corvol, J.C., Czernecki, V., Djebara, M.B., Golmard, J.L., Agid Y., and Hartmann, A. 2010 Segmental progression of early untreated Parkinson disease: a novel approach to clinical rating, *Journal of Neurology, Neurosurgery and Psychiatry*, **81**, 20-25 (doi:10.1136/jnnp.2008.159699)

Singh, N., Pillay, V., Choonara, Y.E. 2007 Advances in the treatment of Parkinson's disease, *Progress in Neurobiology*, **81**, 29-44 (doi:10.1016/j.pneurobio.2006.11.009)

Skodda, S., Rinsche, H., Schlegel, U. 2009 Progression of dysprosody in Parkinson's disease over time – A longitudinal study, *Movement Disorders*, **24** (5), 716-722 (doi: 10.1002/mds.22430)

Sun, X. 2002 Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio, IEEE Signal Processing Society, *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP '02), Orlando, Florida, 2002

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

Titze, I.R. 2000 *Principles of Voice Production*, 2nd edition, National Center for Voice and Speech, Iowa City, USA

Tolosa, E., Craig, C., Santamaria, J., Compta, Y. 2009 Diagnosis and the premotor phase of Parkinson disease, *Neurology*, **72**, 12-20

Tsanas, A., Little, M.A., McSharry, P.E., Ramig, L.O. 2010a Accurate telemonitoring of Parkinson's disease progression using non-invasive speech tests, *IEEE Transactions Biomedical Engineering*, **57**, 884-893 (doi: 10.1109/TBME.2009.2036000)

Tsanas, A., Little, M.A., McSharry, P.E., Ramig, L.O. 2010b 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 (doi: 10.1109/ICASSP.2010.5495554)

von Campenhausen, S., Bornschein B., Wick, R., Bötzel K., Sampaio, C., Poewe W., Oertel, W., Siebert, U., Berger, K., and Dodel, R. 2005 Prevalence and incidence of Parkinson's disease in Europe, *European Neuropsychopharmacology*, **15**, 473-490 (doi:10.1016/j.euroneuro.2005.04.007)

Zou, H., and Hastie, T. 2005 Regularization and variable selection via the elastic net, *Journal of the Royal Statist. Soc., ser. B*, **67**, 301–320 (doi: 10.1111/j.1467-9868.2005.00503.x)