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

Athanasios Tsanas[1,2,*], Max A. Little[1,2,3], Patrick E. McSharry[1,3,4] and Lorraine O. Ramig[5,6]

[1]*Systems Analysis, Modelling and Prediction (SAMP) group, Mathematical Institute and Department of Engineering Science, University of Oxford, Oxford, UK*
[2]*Oxford Centre for Industrial and Applied Mathematics (OCIAM), Mathematical Institute, University of Oxford, Oxford, UK*
[3]*Oxford Centre for Integrative Systems Biology, Department of Physics, University of Oxford, Oxford, UK*
[4]*Smith School of Enterprise and the Environment, University of Oxford, Oxford, UK*
[5]*Speech, Language, and Hearing Science, University of Colorado, Boulder, CO, USA*
[6]*National Center for Voice and Speech, Denver, CO, USA*

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 (approx. 6000 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.

Keywords: nonlinear speech signal processing; nonlinear regression
and classification; Parkinson's disease; telemedicine; Unified Parkinson's
Disease Rating Scale

1. INTRODUCTION

Parkinson's disease (PD) is a common neurodegenerative disorder with prevalence rates exceeding 100/100 000 [1]. Furthermore, it is possible that these statistics underestimate the problem, since an additional 20 per cent of people with Parkinson's (PWP) are not diagnosed [2]. Given that age is the single most important risk factor for PD onset, particularly after age 50 [3], 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 [4]. Speech disorders are among the earliest indicators of PD onset [5], and are reported in about 90 per cent of PWP [6]; moreover 29 per cent of the patients

Unified Parkinson's Disease Rating Scale (UPDRS)

comprises three components and 44 sections in total, each section spans the range 0–4

component 1
mentation, behaviour and mood

four sections (1–4)

includes mentation, thought disorder, depression, and motivation/initiative

component 2
activities of daily living

13 sections (5–17)

ability to complete daily tasks unassisted, e.g. dressing, walking, writing

component 3
motor (motor-UPDRS)

27 sections (18–44)

muscle problems e.g. tremor, rigidity, posture, stability, bradykinesia

section 5: speech—the clinician assesses whether the subject's vocal output is *understandable* during casual discussion

section 18: speech—the clinician assesses whether the subject's vocal output is *expressive* during casual discussion
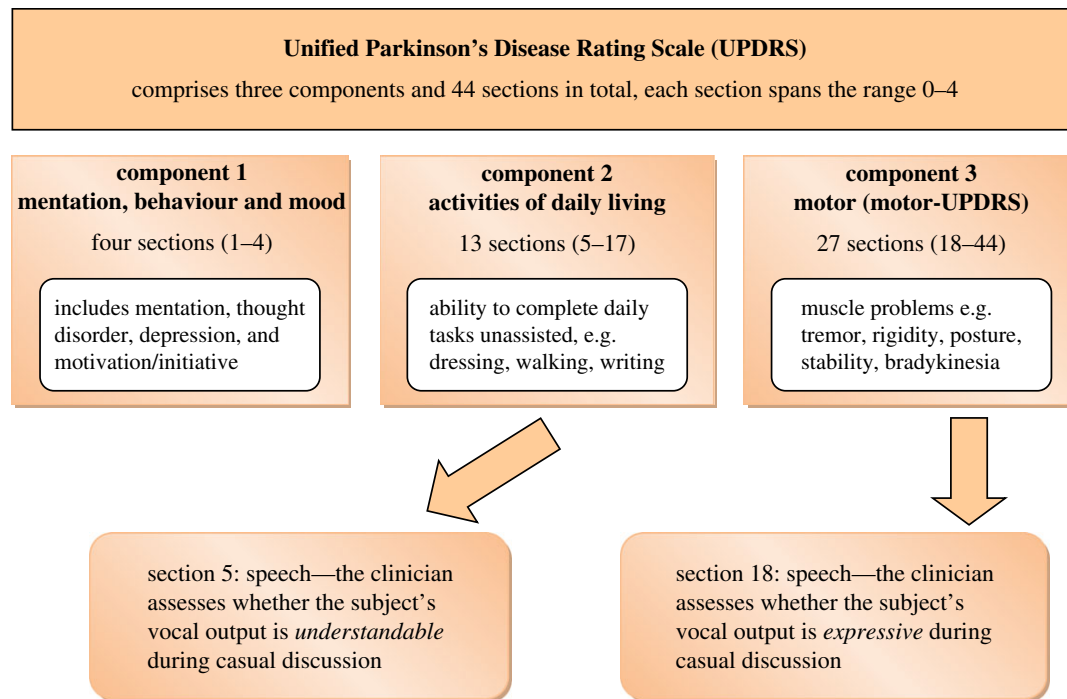
Figure 1. Overview of the clinical metric that quantifies average Parkinson's disease symptom severity, the Unified Parkinson's Disease Rating Scale (UPDRS). Speech appears explicitly twice. (Online version in colour.)

themselves regard speech impairment as one of their most troublesome symptoms [7]. In addition, there is ample empirical evidence for speech degradation as the disease progresses [5,8,9], typically attributed to reduced voice amplitude (*hypophonia*), and increased breathiness (noise) in the PWP's voice [4,6].

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 [10]. However, early diagnosis and frequent disease tracking are critical to maximizing the effect of treatment [4,11]. 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*) [12,13] 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 [13]. 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*: (i) Mentation, behaviour and mood (four sections); (ii) Activities of daily living (13 sections),

assessing whether PWP can complete daily tasks unassisted; and (iii) 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 to 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 and 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 healthcare 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 [14]. 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

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

	males (28 subjects)	females (14 subjects)
age (years)	mean ± s.d.: 64.8 ± 8.1, min. 49, max. 78, median 65	mean ± s.d.: 63.6 ± 11.6, min. 36, max. 85, median 64
weeks since PD diagnosis	mean ± s.d.: 63.0 ± 61.9, min. 1, max. 260, median 48	mean ± s.d.: 89.7 ± 81.2, min. 4, max. 252, median 60
motor-UPDRS (baseline, 3-months, 6-months)	mean ± s.d.: (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 ± s.d.: (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 ± s.d.: (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 ± s.d.: (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)

subjects can elicit steady phonation, whereas subjects with some form of vocal impairment cannot [15]. 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 [16], and is therefore common in general speech clinical practice [15].

Previous studies used speech signals aiming to separate PWP from healthy controls [5,17], and in the past year some authors highlighted the importance of exploring the topic of mapping speech signals to UPDRS [9,14] 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 [18,19]. 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) [12]. 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 [18].

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 [14], recently summarized in [18]. 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.

A schematic 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 [20]. Further details of the AHTD trial can be found in [14].
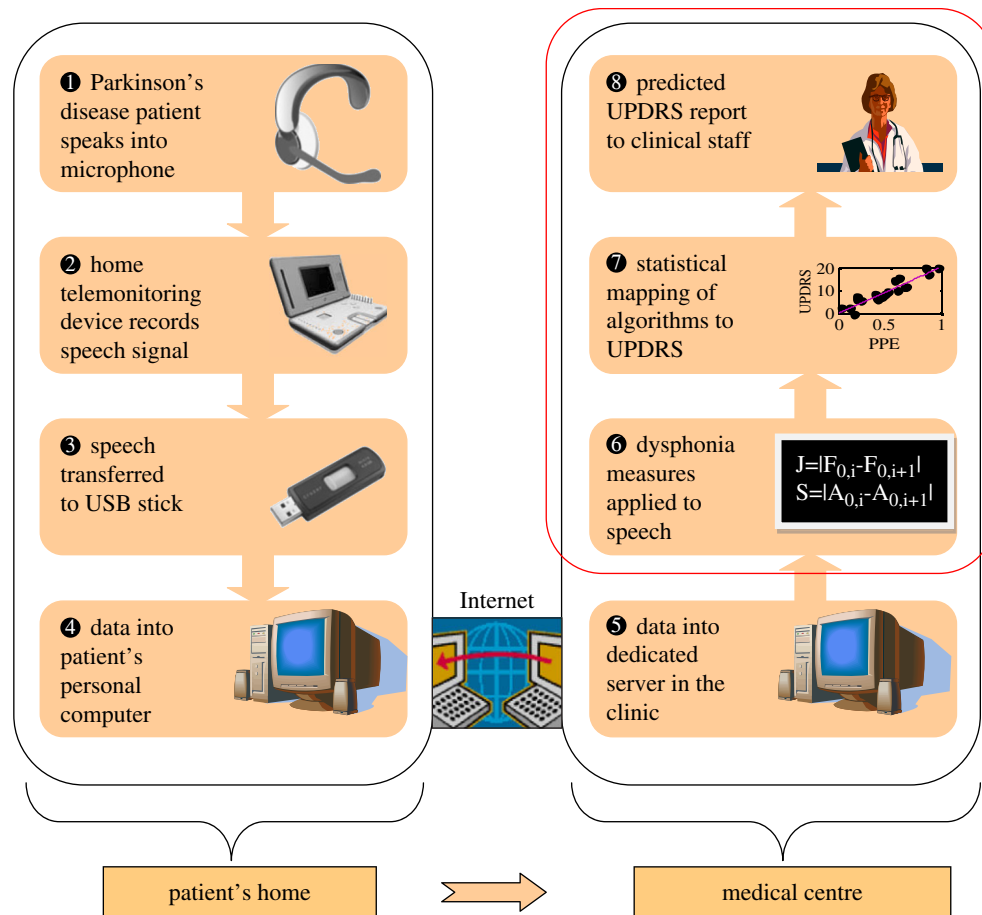
Figure 2. Schematic of the steps from the data acquisition up to UPDRS estimation. The device that collects the data from the Parkinson's disease (PD) patient is known as the at-home-testing-device (AHTD). The black encircled box (steps 6–8) is the focus of this study. (Online version in colour)

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

After initial screening to remove flawed phonations (too short, patient coughing, failure to capture phonation onset), we processed 5875 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: (i) extracting features characterizing the underlying patterns of the speech signals using signal processing algorithms (*feature extraction*),

(ii) selecting a parsimonious subset of these features comprising relevant and minimally overlapping information with regard to UPDRS prediction (*feature selection*), and (iii) 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
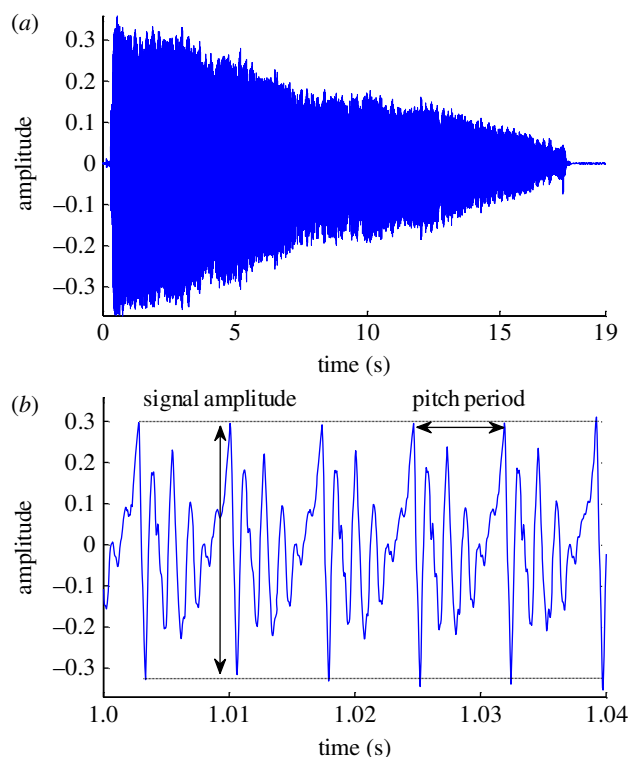
Figure 3. (*a*) Typical sustained vowel phonation signal. (*b*) The same signal magnified in the time axis. The horizontal axes are time in seconds and the vertical axes amplitude (no units). Clear overall amplitude decay over the duration of the phonation can be seen in panel (*a*). A careful look at the magnified signal (*b*) reveals that it is not exactly periodic, a characteristic that many dysphonia measures aim to address. (Online version in colour.)

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 [15]), 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 [21]. In general, people with voice disorders cannot elicit *steady* phonations [15], and speech signal processing algorithms attempt to quantify this inefficiency at converting steady airflow from the lungs into stable voice.

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 5875 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 [22] to extract 13 commonly used measures [18,19] and three new measures we had proposed recently [17,23]. In this study, all algorithms were implemented in MATLAB using the equations described in the electronic supplementary material, §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 5875 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 [24,25]. 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 [18,19]. 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 were non-normal, so we used the non-parametric Spearman's 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's 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 and Y are random variables [26], 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 and 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 §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) [27], and a popular LASSO extension, the *elastic net* [28]. Details of these algorithms and their promising *sparsity-promoting* properties can be found in [27–29]. For both algorithms we computed the entire regularization solution paths [29].

3.4. Regression and classification: mapping dysphonia measures to UPDRS

The analysis in §3.2 provides a 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 $f(x) = y$ which maps the dysphonia measures $x = (x_1 \ldots x_M)$, 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. $y \in R^+$, 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 [18,19], indicating that nonlinear methods may be more appropriate. In particular, we have experimented with *classification and regression trees* (CART), and *random forests* (RFs). Both CART and RF were tested working in both regression and classification modes.

CART was the method of choice in [18] because it has been described as the best off-the-shelf mapping algorithm in supervised learning contexts [29]. 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 into 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 [29].

A natural extension of CART is RFs, 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 [30]. 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 RFs are effective in various prediction tasks, while they do not overfit as more trees are added to the RF [30]. For more information on RF we refer the reader to [29].

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 $5875 \times M$ matrix of feature vectors with all the data (*design matrix*), we used a design matrix of size $4010 \times M$ for male and $1865 \times M$ for female PWP. These design matrices

Table 3. Maximum relevance and correlations of dysphonia measures with UPDRS for males. 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 §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's 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 ($N = 4010$ phonations). The F_0 subscript text refers to the algorithm used to extract it.

measure	description	motor-UPDRS relevance and correlation		total-UPDRS relevance and correlation	
		MI	Spearman's r	MI	Spearman's r
VFER-NSR$_{\text{TKEO}}$	ratio of the sum of the log-transformed mean TKEO of the band-pass signals for frequencies more than 2.5 kHz to the sum of the mean TKEO of the band-pass signals for frequencies less than 2.5 kHz	0.105	0.159	**0.132**	0.187
DFA	characterizes the extent of turbulent noise, quantifying its stochastic self-similarity [23]	0.078	-0.162	**0.115**	-0.205
7th MFCC	7th Mel frequency Cepstral coefficient [33]	0.079	-0.066	**0.108**	0.0070
6th MFCC	6th Mel frequency Cepstral coefficient [33]	0.106	-0.277	**0.102**	-0.294
$F_{0,\text{sun}} - F_{0,\text{expected}}$	mean difference of the cycle-to-cycle F_0 estimate (extracted using Sun's algorithm [34]) 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 [33]	0.090	0.149	**0.099**	0.169
4th MFCC	4th Mel frequency Cepstral coefficient [33]	0.088	-0.082	**0.098**	-0.061
0th MFCC	0th Mel frequency Cepstral coefficient [33]	0.079	0.171	**0.099**	0.197
8th MFCC	8th Mel frequency Cepstral coefficient [33]	0.106	0.276	**0.095**	0.259
8th delta MFCC	8th delta Mel frequency Cepstral coefficient (First derivative of 8th MFCC) [33]	0.073	0.181	**0.093**	0.205

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 (4010 for males and 1865 for females) phonations was split into a training subset of $0.9 \cdot N$ (3609 and 1679) 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. Similarly to our previous work [18,19], we compared model performance on the basis of *mean absolute error* (MAE) for each of the 100 runs for the training and testing subsets:

$$\text{MAE} = \frac{1}{N} \sum_{i \in Q} |\hat{y}_i - y_i|, \quad (3.1)$$

where $\hat{y}_i$ is the predicted UPDRS and y_i the actual UPDRS for the ith entry in the training or testing subset, N the number of phonations in the training or testing subset, and Q contains the indices of that set.

Errors over the 100 cross-validation realizations 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's $r = 0.464$, MI $= 0.153$ for males, and ($p < 0.05$), Spearman's $r = 0.323$, MI $= 0.199$ for females. Similarly, the relationships between speech-UPDRS and total-UPDRS are ($p < 0.001$), Spearman's $r = 0.552$, MI $= 0.22$ for males, and ($p < 0.05$), Spearman's $r = 0.323$, 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,

Table 4. Maximum relevance and correlations of dysphonia measures with UPDRS for females. 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—§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's 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 ($N = 1875$ phonations). The F_0 subscript text refers to the algorithm used to extract it.

measure	description	motor-UPDRS relevance and correlation		total-UPDRS relevance and correlation	
		MI	Spearman's r	MI	Spearman's r
Std $F_{0,\text{mixture}}$	standard deviation of the extracted $F_{0,\text{mixture}}$	0.205	0.475	**0.216**	0.470
Std $F_{0,\text{Rapt}}$	standard deviation of the extracted $F_{0,\text{Rapt}}$	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	0th delta Mel frequency Cepstral coefficient [33]	0.200	-0.327	**0.187**	-0.344
$F_{0,\text{praat}} - F_{0,\text{expect}}$	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	1st delta Mel frequency Cepstral coefficient [33]	0.135	-0.047	**0.170**	-0.031
log energy	estimate of the logarithmic energy [33]	0.179	-0.458	**0.170**	-0.487
$F_{0,\text{mixture}} - F_{0,\text{exp}}$	mean difference of the cycle-to-cycle F_0 estimate (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
$F_{0,\text{Rapt}} - F_{0,\text{expected}}$	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
$\psi(F_0)\text{pret}$	5th percentile of the TKEO of the fundamental frequency values, obtained with the mixture algorithm	0.177	-0.411	**0.153**	-0.369

giving a visual impression of the distribution of the dysphonia signal processing values and their relationship to UPDRS.

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 characterized using higher order moments (the Spearman's correlation coefficient fails to quantify these relationships). Conversely, given two dysphonia measures (for example, the VFER-NSR$_{\text{TKEO}}$ 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$_{\text{TKEO}}$ and UPDRS relies more on higher order moments.

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 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 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.

4.2. Feature selection and statistical mapping of features to UPDRS

As described in §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 study 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 realizations of the 10-fold cross validation. That is, the order of the initially selected features was almost the same across each cross-validation realization used in feature selection. In §2.1 of the electronic supplementary material we compare the 15 most important features selected by the two algorithms.
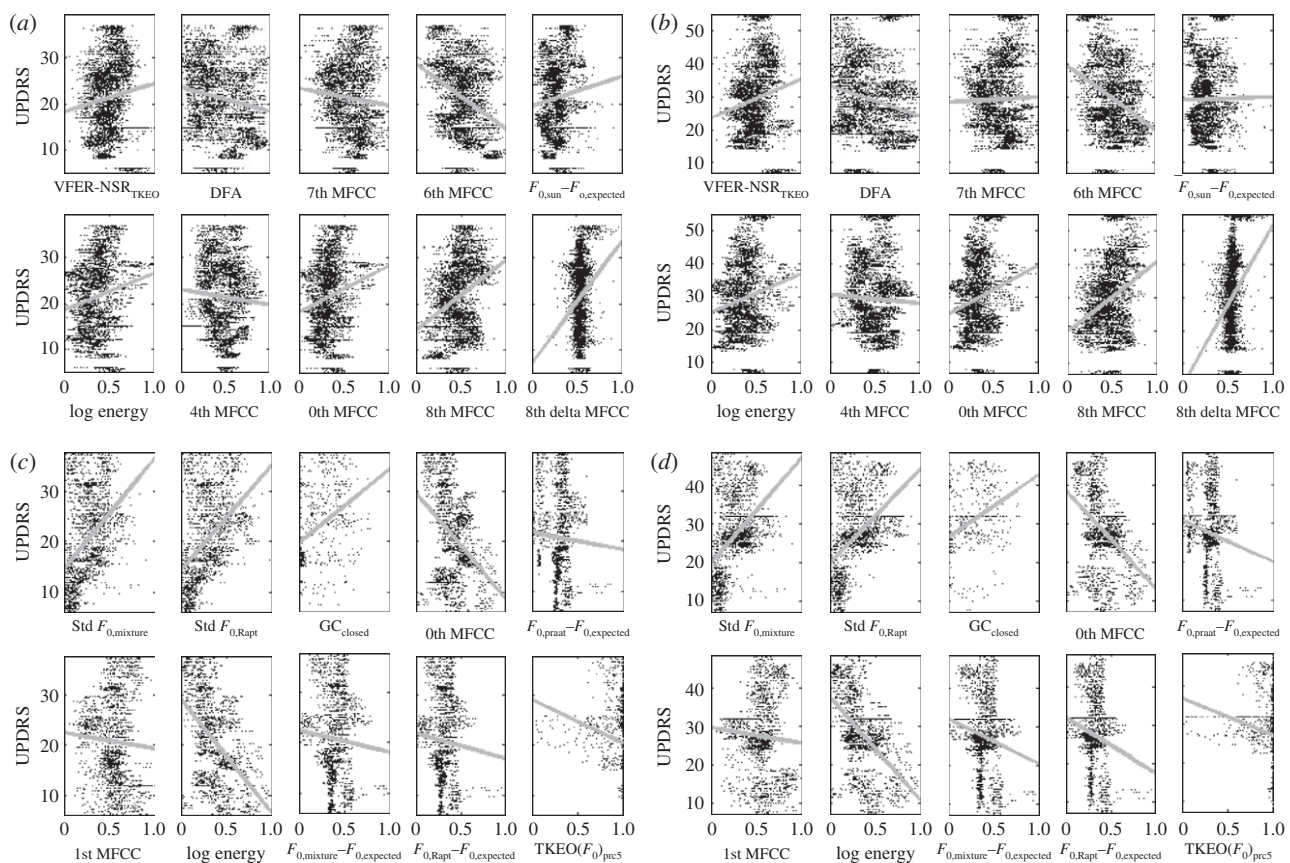
Figure 4. Scatter plots of the most relevant dysphonia measures against motor UPDRS ((*a*) males and (*b*) females) and total UPDRS ((*c*) males and (*d*) females), using the measures presented in tables 3 and 4. The horizontal axes are the normalized dysphonia measures and the vertical axes correspond to UPDRS. The grey lines are the best linear fit obtained using iteratively reweighted least squares—see [35] for details.

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 [29]: 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 [18,19]. The generalization ability of the models is verified by the fact that the in-sample and out-of-sample errors were similarly low.

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 [18,19], 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 [31]. We refer to the electronic supplementary material for details and a comparison of the results using the methodology of this study and [31] in replicating the clinical evaluation

Table 5. Selected dysphonia measure subsets for males and females. 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's 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 and 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 §3.2 for details). Descriptions of the dysphonia measures appear in §1 of the electronic supplementary material.

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

(UPDRS assessment by the clinical rater) of those three elements. Not surprisingly, it appears that accelerometers are better suited compared with 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 §5. 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 six-month duration of the AHTD trial (approx. 140 speech signals × M dysphonia

Table 6. Summary of the mean absolute error (MAE) results of this study, and comparison with the results of previous studies. The reported MAE results were obtained with the random forests (RF) working in classification mode. The errors are reported in the form mean $\pm$ s.d. In [18,19] 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 [12] and the results in this study demonstrate, for the first time, that a machine-learning approach can do better than this benchmark.

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 [18]	5.95 ± 0.19	7.52 ± 0.25
selected feature subset in [19]	6.57 ± 0.16	8.38 ± 0.23

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 (approx. $25 \times M$). However, we have noted 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 with 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 three-month visit and subsequently reduced on the six-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 three-month visit and subsequent sharp decrease in the six-month visit). The female subject in figure 5c,d is the individual we have used previously [18]. Inspection of figure 5c,d and the tracking figure of [18] verifies the superiority of the approach developed in the current study in remotely 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 [18] and [19] where the UPDRS was estimated to within 7.5 points. The improvement in the UPDRS estimation of this study is attributed to two factors: (i) more sophisticated speech signal processing algorithms which uncover novel PD dysphonia patterns, (ii) the use of RFs, 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) [12], 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
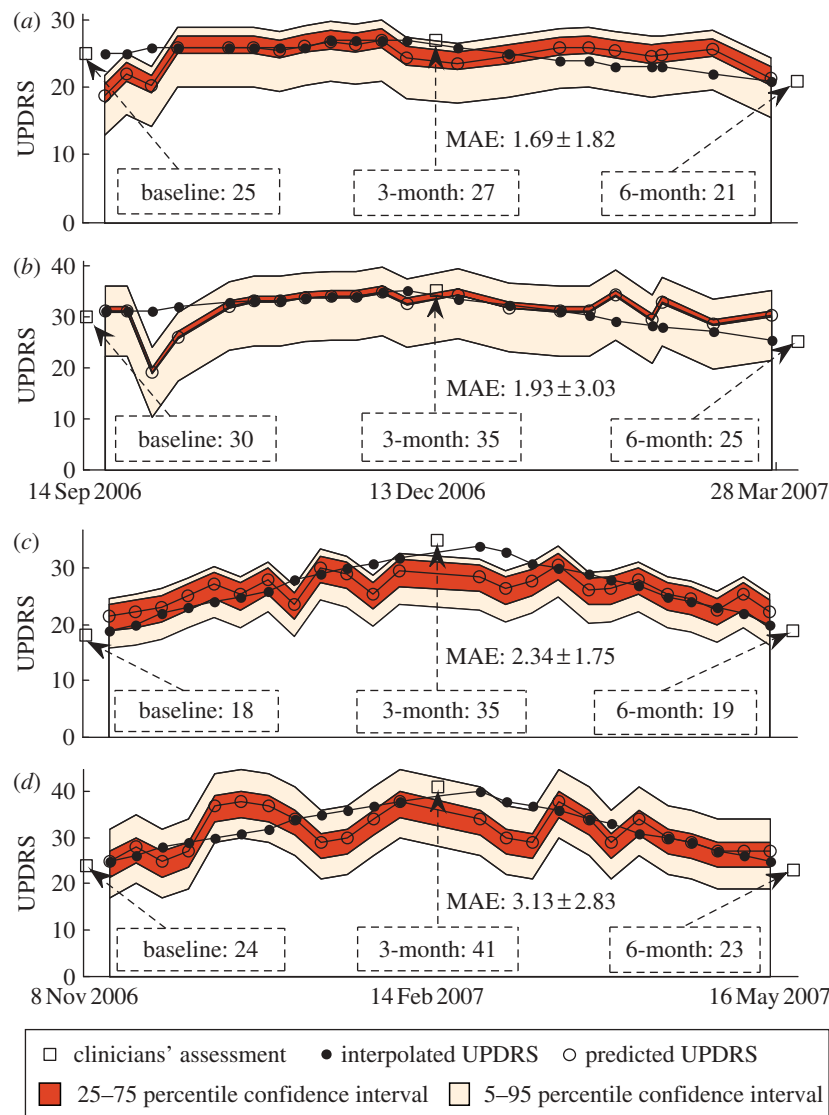
Figure 5. Motor-UPDRS ((*a*) male and (*c*) female subject) and total-UPDRS ((*b*) male and (*d*) female subject) tracking over the six-month trial period for a male and a female subject with irregular UPDRS pattern. The 'baseline', '3-month' and '6-month' UPDRS scores are shown. The out-of-sample MAE 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. (Online version in colour.)

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 probable 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 on 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 [18]. 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 [32], and given that women have higher average F_0 [14], it is plausible that even slight distortions in vocal performance (for example, aperiodic F_0) reflect voice pathology in females with high probability, while 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, non-parametric 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 [30] 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 [18], 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 6000 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 the 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 noted 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.

REFERENCES

1 von Campenhausen, S. *et al.* 2005 Prevalence and incidence of Parkinson's disease in Europe. *Eur. Neuropsychopharmacol.* **15**, 473–490. (doi:10.1016/j.euroneuro.2005.04.007)

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

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

4 Pahwa, R. & Lyons, E. 2007 *Handbook of Parkinson's disease*, 4th edn. New York, NY: Informa Healthcare.

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

6 Ho, A., Iansek, R., Marigliani, C., Bradshaw, J. & Gates, S. 1998 Speech impairment in a large sample of patients with Parkinson's disease. *Behav. Neurol.* **11**, 131–137.

7 Hartelius, L. & Svensson, P. 1994 Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey. *Folia Phoniatr Logop* **46**, 9–17. (doi:10.1159/000266286)

8 Holmes, R. J., Oates, J. M., Phyland, D. J. & 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)

9 Skodda, S., Rinsche, H. & Schlegel, U. 2009 Progression of dysprosody in Parkinson's disease over time—a longitudinal study. *Mov. Disord.* **24**, 716–722. (doi:10.1002/mds.22430)

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

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

12 Post, B., Merkus, M. P., de Bie, R. M. A., de Haan, R. J. & Speelman, J. D. 2005 Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Mov. Disord.* **20**, 1577–1584. (doi:10.1002/mds.20640)

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

14 Goetz, C. G. *et al.* 2009 Testing objective measures of motor impairment in early Parkinson's disease: feasibility study of an at-home testing device. *Mov. Disord.* **24**, 551–556. (doi:10.1002/mds.22379)

15 Titze, I. R. 2000 *Principles of voice production*, 2nd edn. Iowa City, IA: National Center for Voice and Speech.

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

17 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 Trans. Biomed. Eng.* **56**, 1015–1022. (doi:10.1109/TBME-00342-2008)

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

19 Tsanas, A., Little, M. A., McSharry, P. E. & Ramig, L. O. 2010 Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression. *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2010)*, *Dallas, Texas, USA.* pp. 594–597. IEEE Signal Processing Society.

20 Ho, A., Bradshaw, J., Iansek, R. & Alfredson, R. 1999 Speech volume regulation in Parkinson's disease: effects of implicit cues and explicit instructions. *Neuropsychologia* **37**, 1453–1460. (doi:10.1016/S0028-3932(99)00067-6)

21 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 Trans. Biomed. Eng.* **53**, 1943–1953. (doi:10.1109/TBME.2006.871883)

22 Boersma, P. & Weenink, D. 2009 *Praat: doing phonetics by computer (Computer program)*. See http://www.praat.org/.

23 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. *Biomed. Eng. Online* **6**, (doi:10.1186/1475-925X-6-23)

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

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

26 Cover, T. M. & Thomas, J. A. 2006 *Elements of information theory*, 2nd edn. New York, NY: Wiley-interscience.

27 Tibshirani, R. 1996 Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B* **58**, 267–288.

28 Zou, H. & Hastie, T. 2005 Regularization and variable selection via the elastic net. *J. R. Statist. Soc. Ser. B* **67**, 301–320. (doi:10.1111/j.1467-9868.2005.00503.x)

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

30 Breiman, L. 2001 Random forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)

31 Patel, S. *et al.* 2009 Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. *IEEE Trans. Inform. Technol. Biomed.* **13**, 864–873. (doi:10.1109/TITB.2009.2033471)

32 Baken, R. J. & Orlikoff, R. F. 2000 *Clinical measurement of speech and voice*, 2nd edn. San Diego, CA: Singular Thomson Learning.

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

34 Sun, X. 2002 Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002), 2002, Orlando, FL*. IEEE Signal Processing Society

35 Bishop, C. M. 2007 *Pattern recognition and machine learning*, 6th edn. New York, NY: Springer.