Using the cellular mobile telephone network to remotely monitor Parkinson’s disease symptom severity

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

Abstract—Telemonitoring of Parkinson’s Disease (PD) has attracted interest because of the potential to make a lasting, positive impact on the life of patients and their carers. Purposeful-built devices have been developed that record various signals which can be associated with average PD symptom severity, as quantified on standard clinical metrics such as the Unified Parkinson’s Disease Rating Scale (UPDRS). Speech signals are particularly promising in this regard, because they are easy to self-collect without the use of expensive, dedicated hardware. Previous studies have demonstrated replication of UPDRS to within less than 2 points of a clinical raters’ assessment of symptom severity, using high-quality speech signals collected using dedicated telemonitoring hardware. Here, we investigate the potential of using the standard voice-over-GSM (2G) or UMTS (3G) cellular mobile telephone networks for PD telemonitoring, networks that, together, have greater than 5 billion subscribers worldwide. We test the robustness of this approach using a simulated noisy mobile communication network over which speech signals are transmitted, and approximately 6000 recordings from 42 PD subjects. We show that UPDRS can be estimated to within about 3.5 points difference from the clinicians’ assessment, which is clinically useful given that the inter-rater variability for UPDRS can be as high as 4-5 UPDRS points. This provides compelling evidence that the existing voice telephone network is more than adequate for inexpensive, mass-scale PD symptom telemonitoring applications.

Index Terms—Decision support tool, Parkinson’s disease, mobile phones, nonlinear speech signal processing, telemedicine

I. INTRODUCTION

PARKINSON’S disease (PD) is a chronic neurodegenerative disorder characterized by the progressive deterioration of motor function as well as the emergence of considerable non-motor problems [1]. The PD incidence rate is approximately 20/100,000 [2] and the prevalence rate exceeds 100/100,000 [3]; moreover it is believed that an additional 20% of people with Parkinson’s (PWP) might be undiagnosed [4]. Early PD stages are mainly characterized by three hallmark symptoms: bradykinesia (slow and reduced amplitude of movement), rigidity (resistance to passive movement), and tremor (while at rest) [5]. Speech disorders, which are of particular interest in this study, may be amongst the earliest PD onset indicators [6], and are reported in the vast majority of PWP [7]. Furthermore, strong empirical evidence has emerged linking speech performance degradation and PD symptom severity [6], [8], [9].

PD onset was believed to be due to dopaminergic neuron reduction in the basal ganglia, but it is now recognized that PD is also characterised by the degeneration of numerous non-dopaminergic pathways [1]. Medication and surgical intervention can alleviate some of the symptoms and improve quality of life for most PWP [10], although there is no known cure currently. To optimize treatment, PWP are typically followed up by expert clinical staff at regular (three to six month) intervals. More frequent monitoring of PD progression would be of considerable benefit, for example, to optimize treatment regimes, but this is not possible given the available resources and the established assessment setting which requires the physical presence of PWP in the clinic.

In current clinical practice, medical raters physically examine PWP and map symptom severity on appropriate clinical scales (metrics). The Unified Parkinson’s Disease Rating Scale (UPDRS) [11] is the most widely used clinical metric for quantifying PD impairment [12], and attempts to quantify the full breadth of possible motor (muscle), non-motor PD symptoms, and complications of dopamine replacement therapies. It consists of 42 items, some with subdivisions (that is, some items have two or more entries). We
shall refer to each individual UPDRS entry as a section in this study. The UPDRS is organized around four major components, where each component is composed of a number of sections. The four components are: (1) Mentation, behavior and mood (4 sections, sections 1-4); (2) Activities of daily living (13 sections, sections 5-17), assessing whether PWP can complete daily tasks unassisted; (3) Motor (27 sections, sections 18-44), addressing muscular control; and (4) Complication of therapy (11 sections, sections 45-55), which expresses dyskinesia and disability problems associated with PD treatment. The third component is commonly referred to as motor-UPDRS, and is highly correlated with the total UPDRS value [13]. For untreated PWP the fourth component is not used. The motor-UPDRS ranges between 0 and 108, where 0 denotes healthy state, and 108 severe disabilities. The total UPDRS value is computed by summing up the constituent sections in the components one to three (sections 1-44), and ranges between 0 and 176. In addition to UPDRS, the Hoehn and Yahr (H&Y) scale is a popular clinical metric which provides overall PD stage assessment; recent studies have shown that it is possible to map UPDRS onto H&Y [14]. Other metrics are occasionally used in clinical practice, but here we focus exclusively on UPDRS.

Telemonitoring in healthcare is an emerging field which has the potential to revolutionize contemporary clinical practice. It relies on the use of modern equipment, some of which was not widely available just a decade ago (such as mobile phones and the Internet), in order to facilitate the flow of information between patient and medical expert. This flow enables fast, frequent, remote tracking of disease status, minimizing the need for frequent and often inconvenient physical visits to the clinic, and allowing rapid response to the changing circumstances of patients. In addition, telemonitoring alleviates the workload of medical personnel and the associated financial strains on national health systems. Telemonitoring of PD has attracted considerable interest lately [13], [15], [16], [17], [18]. However, all those studies rely on the use of expensive, specialized hardware-purpose-built to record signals which are characteristic of PD symptoms.

Recently, we demonstrated the considerable potential of speech in this application [13], [18], [19], [20], using high quality speech signals collected with Intel Corporation’s At-Home Testing Device (AHTD) [15]. This device collects high quality speech signals sampled at 24 kHz, following the established recommendation that a sampling frequency of at least 20 kHz should be used to extract clinically useful information [21]. In this study, we investigate whether it is possible to accurately infer UPDRS using speech signals transmitted over the standard cellular mobile voice telephony network, simulating all the major steps involved in this process. The rationale for using the existing voice mobile phone network over specialized, purpose-built hardware is that (a) the existing voice network reaches nearly 75% of the global population, (b) economies of scale and global market competition has brought the price of access down so that it is affordable to a majority of the global population, (c) mobile telephony allows freedom of movement for PWP, eliminating the need to carry additional equipment when leaving home. Data-mining of speech signals obtained using the public telephone network for clinically useful information has recently shown promising results [22], [23]. Similarly, Saenz-Lechon et al. [24] investigated the effect of different data transmission rates in automatic voice pathology detection, and concluded that compressing signals (down to at most 64 kbps) does not prevent accurate detection of pathologies.

We demonstrate that mobile phone technology could indeed be useful in telemonitoring average PD symptom severity further endorsing our previous findings that speech may offer a convenient framework [13], [18], [20].

II. Data

We use the voice data collected by Goetz et al. [15], described in detail in Tsanas et al. [13]. In summary, 52 subjects with idiopathic PD diagnosis up to five years from the time of the baseline clinical visit were recruited into a clinical trial to investigate the potential of the AHTD. All subjects gave written informed consent, and did not receive PD-related treatment for the six-month duration of the trial. They were asked to complete a range of tests weekly during a convenient, pre-specified time window (all tests can be completed in about 20-30 minutes). Sustained vowel “ahh…” phonations, where the subject is asked to sustain vowel phonation at a comfortable pitch for as long and as steadily as possible, were part of the test protocol. Here we focus exclusively on these sustained phonations. Subjects were diagnosed with PD if they had at least two of the three hallmark PD symptoms (bradykinesia, rigidity, tremor), without evidence of other forms of Parkinsonism. We did not apply any exclusion criteria related to specific PD symptoms (e.g. depression). We disregarded data from 10 recruits – two that dropped out of the study early, and from eight additional PWP recruits that did not complete at least 20 valid study sessions during the trial period. Therefore, in this study we analyze data from 42 PWP.

Previously, we demonstrated that partitioning the data by gender is important in this application [13], [20], and hence males and females are studied separately here as well. The 28 male subjects were 64.8±8.1 (mean ± standard deviation) years old, with a PD diagnosis 63.0±61.9 weeks since diagnosis at trial baseline. Their motor-UPDRS scores were: baseline 20.3±8.5, three months into the trial 21.9±8.7, six months into the trial 22.0±9.2, and total-UPDRS scores were: baseline 27.5±11.6, three months into the trial 30.4±11.8, and six months into the trial 31.0±12.4. The 14 female subjects were 63.6±11.6 years old, with a PD diagnosis 89.7±81.2 weeks since diagnosis at trial baseline. Their motor-UPDRS was: baseline 17.6±7.4, three months into the trial 21.2±10.5, six months into the trial 20.1±9.4, and their total-UPDRS was: baseline 24.2±9.1, three months into the trial 27.4±12.1, and six months into the trial 26.8±10.8.

Six sustained vowel “ahh…” phonations were recorded each time the PD subject took the test: four at comfortable level of pitch and loudness, and two at twice the comfortable loudness (elicited with the instruction “twice as loud as the first time”). The signals were sampled at 24 kHz at 16 bit
resolution. After initial processing to remove faulty phonations (e.g. patient coughing, failure to record phonation), we processed 4010 phonations for the male subjects, and 1865 phonations for the female subjects.

Although the phonations were recorded weekly, the actual clinical assessments for motor-UPDRS and total-UPDRS were obtained at trial baseline, three months into the trial, and at six months into the trial. To obtain weekly UPDRS estimates to associate with the phonations we suggested using piecewise linear interpolation going exactly through the measured baseline, three-month and six-month UPDRS assessments [18], [13], [19], [20], [25]. This assertion builds on strong empirical evidence suggesting that average symptom progression in early PD stages (up to about five years) is almost linear in non-medicated patients as observed in clinical metrics [26], [27]. The PWP in the AHTD trial were in early PD stages (up to five years from disease diagnosis) and remained non-medicating for the duration of the trial, aspects which justify the use of piecewise linear interpolation when filling in missing data. The tacit assumption is that PD symptom severity did not fluctuate wildly within the intervals where the clinical scores were obtained. Finally, the interpolated UPDRS values are rounded to the nearest integer.

For further details about the dataset and the AHTD data acquisition hardware, please refer to Tsanas et al. [13].

III. METHODS

A. Simulation of the cellular mobile telephony network

Creating a realistic simulation of the cellular voice telephony network requires the following steps: (a) encoding the AHTD speech signals into bit-streams for transmission, (b) simulate the transmitter, radio channel, and receiver, and (c) decode the transmitted bit-streams back into intelligible speech recordings. It is important to note that this application requires only one way (simplex) communication: PWP call into an automated voice messaging service and leave sustained vowel phonations. Predictions of symptom severity are extracted from these voice messages and clinical personnel suggest the appropriate course of action offline as a result of the estimated UPDRS. Moreover, the sustained vowel phonations need only be a few seconds long, that is, considerably shorter in duration than most telephone conversations. The schematic diagram of the communication system used in this study is presented in Fig. 1.

Speech coding

The original AHTD signals were sampled at 24 kHz, 16 bits. Each speech signal needs to be coded appropriately for transmission over the communication network. 2G (GSM) and 3G (UMTS) cellular networks make use of the AMR-Narrowband codec. This is an Enhanced Full Rate (EFR) algebraic code-excited linear prediction (ACELP) algorithm, with 16-bit algebraic codebook and up to 4 non-zero (pitch) pulses per 160-sample frame. This compresses speech at 8 kHz, 16 bits (64 kbps) into a 12.2 kbps bit-stream. It uses perceptually-weighting to mask the linear prediction residual encoding noise under the formant envelop.

We refer to Chu [28] for further details about this speech codec. This AMR-Narrowband codec provides a high compression ratio (around 10:1) and very good intelligibility for spoken speech in non-noisy environments.

Digital communication network: transmitter, channel, and receiver

The main components of a digital communication system are the transmitter, the channel (physical medium connecting the transmitter and the receiver), and the receiver. The transmitter aims to assist the receiver to correctly recover the speech signal which may be distorted by the channel. Here, we very briefly explain this process and refer to Proakis for a comprehensive treatment [29]. We follow closely the studies of Tsanas [30] and Ampeliotis and Berberidis [31] for the practical implementation.

After speech coding, the bit-stream is fed into the encoder. The transmission takes place in bursts, i.e. segments of the binary sequence. The GSM standard uses 156.25 bits, which includes a 26 bit training sequence to estimate the channel. Here, to illustrate the proposed methodology we used 768 bits in each burst since the channel is assumed to be known to the receiver (very common assumption in the literature, e.g. [31]). The encoder inserts a controlled number of redundant bits to protect the binary data sequence from noise. We used a recursive systematic convolutional encoder with generator matrix \( G(D) = [1((1 + D^2)/(1 + D + D^2))] \) (this error correction method is used to relate bits adding additional redundant bits and hence maximize the probability of correct bit detection in the receiver), with rate \( R=1/2 \) (that is, for every bit of the actual signal, an additional redundant bit is introduced). The resulting bit sequence is fed into the interleaver which re-arranges (permutes) the coded bits, to decorrelate errors between adjacent symbols in the channel. Finally, the mapper takes blocks of the interleaved coded bits and transforms them into symbols suitable for transmission. Here, we used Gray-coding Quaternary Phase Shift Keying (QPSK) which uses two bits per symbol. The signal to noise ratio for the signal transmission was set to 10 dB.

The symbols per burst are transmitted over a channel which introduces additive white Gaussian noise and inter-symbol interference (ISI) which distort the transmitted signal. We assume that the channel is frequency selective and constant during each burst transmission, and also that it is known to the receiver (in practice the channel would need to be estimated using a training sequence embedded in the transmitted signal, which would be known a priori to the receiver). The channel is approximated by an equivalent discrete time linear filter. We used the standard Proakis C channel [29] which severely distorts the transmitted signal.

The main components of a traditional receiver are the equalizer, de-mapper, de-interleaver, and decoder. The aim of the receiver is to accurately recover the original bit-stream. The equalizer exploits the structure of the transmitted symbol constellation to provide a symbol estimate, redressing the
detrimental effects of the channel distortion. In this study, we use an iterative turbo equalizer, which involves a feedback loop from decoder to equalizer to provide prior information about the received bits [32]. We used Tuchler’s Minimum Mean Square Error (MMSE) equalizer as part of the turbo equalization scheme [33]. The de-mapper decomposes the received symbols into bits, the de-interleaver restores the initial order of bits (reversing the bit permutation that took place in the interleaver), and the decoder makes use of the added redundancy of the convolutional encoder in order to make an accurate estimation of the transmitted bits. We used a standard Maximum A Posteriori (MAP) decoder to decipher the transmitted sequence. Turbo-equalizers are characterized by iterating soft information (i.e. bit probabilities) between the equalizer and the decoder; we used five iterations before the final decision on the recovered bits. Finally, the output bits from the decoder are fed into the speech decoder, which reconstructs the digital speech signal, undoing the effect of the speech coding.

B. Methodology to analyze the speech signals recovered at the receiver

We follow three steps to process the recovered phonations and extract clinically useful information: (a) feature extraction, where we apply speech signal processing algorithms to characterize the phonations and extract characteristic patterns (features), (b) feature selection, where a parsimonious (small, information-rich) subset of the originally computed features is selected in order to provide maximally useful information for predicting UPDRS, and (c) feature mapping, where a standard supervised learning algorithm is used to determine a functional form associating the selected feature subset with the clinical outcome (UPDRS). The rationale behind this methodology is that characteristic acoustic patterns in PWP’s voice are indicative of UPDRS. Although confounding factors may affect vocal performance (such as the subject’s emotional state, or some pathological condition not related to PD), it is unlikely these contaminate more than a handful of the approximately 6000 recordings used here.

Feature extraction

We apply the dysphonia measures rigorously defined in Tsanas et al. [13] to the speech signals recovered at the receiver. We refer to that paper for detailed description of the concepts and rationale behind each algorithm. Here, we briefly describe the most important families of dysphonia measures used in this and other studies.

Some of the most widely used dysphonia measures are jitter and shimmer [21], [34]. They seek to capture the physiological observation that the vocal fold vibration pattern is nearly periodic in healthy voices, whilst it is disturbed in pathological voices [34]. Jitter characterizes deviations in fundamental frequency (F0), whereas shimmer characterizes deviations in amplitude. There is no unique definition of those dysphonia measures, and we investigated many jitter and shimmer variants [13] which are algorithmic variations of the same underlying concept. Quantifying vocal fold departure from near periodicity has inspired the development of the Recurrence Period Density Entropy (RPDE) [35], the Pitch Period Entropy (PPE) [36], the Glottis Quotient (GQ) [13], and F0-related measures [13]. GQ can be seen as an improved jitter-like family of measures, but working directly with vocal fold cycles instead of pre-specified segments (e.g. 10 ms) of the speech signal. RPDE expresses the uncertainty in vocal fold cycle duration. PPE quantifies the impaired control of F0 in sustained phonations, taking into account normal vibrato. The F0-related measures include statistical summaries of F0 distributions, and F0 differences compared to average age- and gender-matched healthy controls in the population.

The second group of dysphonia measures characterize signal to noise ratio (SNR)-like quantities. The physiological motivation for this group is that incomplete vocal fold closure leads to the creation of aerodynamic vortices which result in increased acoustic noise. Harmonic to Noise Ratio (HNR) [34], Detrended Fluctuation Analysis (DFA) [35], Glottal to Noise Excitation (GNE) [37], Vocal Fold Excitation Ratio (VFER) [13], and Empirical Mode Decomposition Excitation Ratio (EMD-ER) [13] are typical examples. GNE and VFER analyze the frequency ranges of the signal in bands of 500 Hz. Empirically, we found that frequencies below 2.5 kHz can be treated as ‘signal’, and everything above 2.5 kHz can be treated as ‘noise’ [13] to define SNR measures using energy, nonlinear energy (Teager-Kaiser energy operator) and entropy concepts. EMD-ER is similarly motivated: the Hilbert-Huang transform [38] decomposes the original signal into its constituent components in decreasing order of contributing
frequency. Then, the top (high frequency) components are taken to constitute noise, and the lower frequency components to constitute signal, to obtain SNR-like measures.

Lastly, Mel Frequency Cepstral Coefficients (MFCC) have been traditionally used in speaker recognition applications, but also appear promising in biomedical speech signal processing contexts [13], [39], [40]. MFCCs detect subtle changes in the motion of the articulators (tongue, lips) which are known to be affected in PD [7].

Overall, we applied 132 dysphonia measures to the speech database, each dysphonia measure producing a single real value per voice sample, resulting in a design matrix of size 4010×132 for male PWP and a matrix of size 1865×132 for female PWP. There were no missing entries in either design matrix.

Feature selection

The use of a large number of features (132 in this study) makes it extremely difficult to discern meaningful patterns in the data, and may often be detrimental in the process of mapping the features onto the clinical outcome UPDRS. This problem is known as the curse of dimensionality, and arises because adequate population of the feature space requires the number of voice samples increases exponentially with the number of features [41]. Contemporary algorithms that can map features onto outcomes may be very robust to the inclusion of potentially noisy or irrelevant features, and their predictive power may or may not be severely affected; however, a smaller feature set always facilitates insight into the problem by allowing interpretation of the most predictive features [42], [43]. An exhaustive search through all possible feature subset combinations is computationally impractical; feature selection (FS) algorithms are a principled approach to selecting a smaller (lower dimensional) feature subset. We refer to Guyon et al. [43] for a detailed overview of feature selection.

Here, we compared four FS algorithms: (1) Least Absolute Shrinkage and Selection Operator (LASSO) [44], (2) Minimum Redundancy Maximum Relevance (mRMR) [45], (3) RELIEF [46], and (4) feature importance in Random Forests (RF) [47]. LASSO introduces the L1-norm penalty into classical linear regression, penalizing the absolute value of the coefficients; sweeping through the regularization parameter leads some regression coefficients to shrink to zero, which effectively indicates that the associated features are irrelevant and can be removed. The LASSO has oracle properties (correctly identifying all ‘true’ predictive features) in sparse settings with low correlations amongst features, but if those conditions are violated, then non-predictive features may be selected [48]. The mRMR approach relies on a heuristic criterion to compromise between relevance (association strength of features with the outcome) and redundancy (association strength between pairs of features). It is a greedy algorithm (incrementally adding one feature at a time), and takes into account only pairwise redundancies neglecting complementarity (joint association of features towards predicting the response). RELIEF is a margin maximization algorithm [49], selecting features that contribute to the separation of samples from different classes. Inherently, RELIEF uses complementarity as part of the FS process, but does not account for feature redundancy. SIMBA is an extension of RELIEF, with a similar conceptual basis. Its advantage over RELIEF is that it reweights the samples as part of an updating algorithm which may lead to discarding redundant features. We defer discussion of the feature importance in RF to the next subsection. All these feature selection algorithms have demonstrated encouraging results in statistical machine learning contexts over a wide range of applications.

We apply the methodology for FS selection that was previously described in Tsanas et al. [40], which is provided here again for completeness. The feature subsets were selected using a cross-validation (CV) scheme, where at each CV iteration only the training data was used. We used 10 CV iterations, where at each iteration the M features (M = 132) for each FS algorithm are selected in descending order. Theoretically, this feature order should be identical for all 10 CV iterations, but often, in practice it is not. Therefore, we apply the following strategy to select the features that appeared most often under each of the FS algorithms. The ultimate aim is to determine one subset for each FS algorithm. For each FS algorithm we build an empty set S which will contain the indices of the features selected. Feature indices are incrementally included in S, one index at each step. For each step K (K is a scalar with value 1 ... M) we find the indices corresponding to the features selected in the 1 ... K search steps for the 10 CV iterations. The index which appears most frequently amongst the 10×K elements that is also not already included in S, is now included as the Kth element in S. Ties are resolved by including the lowest index number. We repeat this process for each of the FS algorithms. Contrary to the other FS algorithms studied here, LASSO may remove features which were selected in prior steps in its incremental search. Thus, prior to the voting scheme described above, we repeated the 10-fold CV process independently for each Kth step in LASSO, in order to obtain the best K features.

The final feature subset S selected for each FS algorithm was used in the subsequent statistical mapping phase.

Feature mapping

In the preceding steps we have computed 132 characteristic patterns from the sustained vowel phonations, and subsequently applied feature selection techniques to obtain subsets of those features. Here, we aim to determine the functional relationship f(X) = y, which maps the dysphonia measures X = (x1 ... xM), where M is the number of features, to the outcome (response) y (motor-UPDRS and total-UPDRS in this study). That is, we want to obtain a classifier that will use the dysphonia measures to accurately predict UPDRS. There is a large literature on supervised classification, and we refer to Hastie et al. [42], and Bishop [41] for a broad overview of this area. Here, we experimented with two powerful classifiers: Random Forests (RF), and Support Vector Machines (SVM).

RF is an intuitively appealing ensemble classifier, where a large number of decision trees (each of those decision trees is an individual base classifier with a prediction function f) vote for the most frequent class. RF has two free parameters: (a)
the number of decision trees (which should be fairly large; we used 500 decision trees), and (b) the number of features over which to search to construct each branch of each tree. Breiman [47] has shown that RF does not overfit with an increasing number of decision trees included in the ensemble, and also that it is fairly insensitive to the choice of the other free parameter. Following the suggestion of Breiman [47], we used the square root of the number of input features over which to search to construct each branch of each decision tree, and also experimented with doubling and halving this value. Another attractive aspect of RF is determining feature importance as an inherent part of search strategy for building the individual decision trees. Feature importance can be thought of as a weight associated with each feature, which is related to the use of that particular feature to split the data into separate regions, at the nodes of each decision tree. Hence, the feature importance concept accounts for relevance, redundancy, and complementarity, and has shown promising results in detecting parsimonious datasets [50]. We refer to Breiman [47] for further details about the specifics of RF and feature importance.

SVMs were originally proposed for binary classification problems, and were subsequently extended to multi-class classification problems. In their basic form for binary classification, they aim to construct an optimal separating hyper-plane in the feature space, maximizing a geometric margin between samples from the two classes. In practice, typical data cannot be linearly separated; hence SVMs transform the data into a higher dimensional space using the kernel trick, and construct the separating hyperplane in that space [42]. Contrary to RF, SVMs are known to be particularly sensitive to the values of their free parameters [42], which vary depending on the choice of kernel. In this study, we used the LIBSVM implementation [51] and followed the suggestions of its developers: we linearly scaled each of the input features to lie in the range [-1, 1], and used a Gaussian radial basis function kernel (which has two free parameters requiring optimization). The determination of the optimal values of the kernel parameter \( \gamma \) and the penalty parameter \( C \) was decided using a grid search of possible values. We selected the pair \((C, \gamma)\) that gave the lowest CV error metric (see the following section for details). Specifically, we searched over the grid \([C, \gamma]\), where \( C = [2^{-5}, 2^{-3}, ..., 2^5] \), and \( \gamma = [2^{-5}, 2^{-3}, ..., 2^3] \). Once the optimal parameters \( C \) and \( \gamma \) were determined, we trained and tested the SVM using these parameters.

Cross-validation and model generalization

As in previous studies [13], [18], [20] we used 10-fold CV to test the generalization performance of the RF and SVMs. Conceptually, CV provides a good estimate of the accuracy with which UPDRS may be predicted on a new dataset using the classifier described above, assuming the new dataset has similar statistical characteristics to the data used to train and test the classifier. Specifically, we split the initial dataset comprising \( N \) (4010 for males and 1865 for females) phonations into a training (in sample) 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. For statistical confidence, the process was repeated a total of 100 times, randomly permuting the data each time before splitting into training and testing subsets. As in previous studies [13], [18], [19], [20], we used mean absolute error (MAE) to assess the model performance:

\[
\text{MAE} = \frac{1}{N} \sum_{i \in Q} |\hat{y}_i - y_i| \tag{1}
\]

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 CV realizations were averaged.

IV. Results

We computed correlation coefficients and normalized mutual information [13] to quantify the strength of statistical association of the features with the outcome variable, and then to compare this association strength to previous findings [13] (results not shown due to space constraints). We have found that, as expected, the features in the present study had lower association strength with motor-UPDRS and total-UPDRS than in earlier studies that used full bandwidth speech [13]. We report the accuracy with which UPDRS can be predicted in Table I for males, and Table II for females. For each FS algorithm, the final number of features \( K \) is determined using the one standard error rule [42]: adhering to the principle of parsimony, we fix \( K \) to be the number of features where MAE is up to one standard deviation larger than the globally lowest MAE obtained with the feature subsets from that FS algorithm.

The MAE for motor-UPDRS is 2.91 \( \pm \) 0.23 for males and 2.38 \( \pm \) 0.23 for females, whilst the MAE for total-UPDRS is 3.43 \( \pm \) 0.27 for males and 2.91 \( \pm \) 0.27 for females. By comparison, in a recent study in this application where high-quality, high-bandwidth, uncompressed speech signals from the AHTD were used instead, the MAE reported for total-UPDRS was [20]: 1.49 \( \pm \) 0.14 for males, and 2.14 \( \pm \) 0.25 for females. Overall, these results suggest that there is indeed clinically important information in the higher frequency ranges (that is discarded in the GSM codec), and this suggestion concurs with that of clinical speech specialists [21]. However, it seems that clinically useful information can still be extracted from much lower-quality, low-bandwidth, compressed speech signals.

V. Discussion

We have extended previous findings which demonstrated that speech signals may be promising in telemonitoring for PD, by investigating the robustness of the GSM mobile telephone network for this application. We found strong evidence that the existing GSM network, which to-date reaches more than 5 billion subscribers, enables clinically accurate UPDRS estimation.

In Tsanas et al. [20], where the high-quality signals
obtained from Intel’s AHTD were used, we reported that UPDRS could be estimated to within 1.5 UPDRS points for males and 2.2 UPDRS points for females; here we demonstrated that UPDRS can be estimated to within approximately 4.1 UPDRS points for males, and 3.8 UPDRS points for females. We argue that this loss in accuracy of UPDRS, which is due to bandwidth restriction and/or channel transmission error is acceptable in practice because most PWP who could benefit from remote symptom tracking, are unlikely to have access to expensive, dedicated hardware such as the AHTD. We emphasize that the accuracy with which UPDRS is estimated even in this scenario of restricted quality speech, is very close to the inter-rater variability (difference in UPDRS score between two expert clinicians), which is about 4.5 points [52]. Therefore, speech over GSM remains clinically useful here, and could be used as a decision support tool to aid clinicians in remote, non-invasive PD symptom severity tracking. The automatic assessment of voice pathologies using signals transmitted over the public telephone network was recently shown to be promising in related applications [22], [23].

Concurring with previous findings, we have found that a parsimonious speech feature subset actually improves the out-of-sample MAE, and is also more amenable to interpretation [13], [18], [20]. Random Forests outperformed SVMs consistently and significantly ($p<0.001$), although we cannot provide a clear theoretical justification for this finding. As indicated in previous studies [40], more detailed empirical and theoretical analysis is required to understand which classification algorithm is likely to lead to more accurate prediction for similar datasets.

The non-classical dysphonia measures (mainly VFER and MFCCs) are consistently selected as the most predictive features by the feature selection algorithms. One very interesting new finding in this study is that UPDRS estimation in males deteriorates considerably more compared to UPDRS estimation in females. We emphasize that this loss in accuracy of UPDRS estimation in males deteriorates considerably more compared to UPDRS estimation in females. We argue that this loss in accuracy of UPDRS estimation in males deteriorates considerably more compared to UPDRS estimation in females.
estimation in females as a result of the lower quality speech signals. This may be related to the bandwidth restriction, but may also be a consequence of the finite bit allocation available to reproducing the pitch period with pitch pulses in the ACELP codec. It could also be due to the increased noise that is masked by the formants in the perceptually-weighted linear prediction: this noise may not be heard, but may, nonetheless, be important in PD.

The results of this study confirm the established view in the clinical speech community suggesting that speech signals of at least 20 kHz should be preferred in clinically applications because there is useful information in the higher frequencies of the spectrum [21]. Nevertheless, the performance degradation as a result of the use of the lower-quality GSM coding and communication framework is unlikely to be prohibitive for clinically useful UPDRS prediction. We conjecture that this may also be the case for other voice pathologies. We hypothesize that the speech community may have, hitherto, been overly pessimistic in the need for very high quality speech signals [21] in clinical speech science.

We stress that the results reported in this study were obtained in a simulated digital communications framework involving the GSM standard. Additional tests in real-world contexts using actual mobile phones would be required to validate the robustness of the presented methodology. For example, we have not simulated the effect of drop-outs due to cell handoff, or switching between 2G/3G, or quality reduction due to the user not placing the phone close to their mouth. For this reason, our channel is chosen to be extremely noisy which introduces quite severe speech signal quality degradation. Additional factors that are hard to control, such as the frequency response of the mobile phone’s microphone, might need to be taken into account in a real application.

Telemonitoring in healthcare has received considerable attention lately, but global adoption is always constrained by the prohibitive costs associated with specialized telemonitoring hardware. The exploration of highly cost-effective solutions, such as exploitation of existing cellular or PSTN telephone networks may be a critical step towards more widespread diffusion of this promising technology. We envisage the results of this study being a first step towards practical, affordable, and accurate telemonitoring of PD for the population at large.

ACKNOWLEDGMENT

We 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, Ejaž A. Shamim, Helen Bronte-Stewart, Jennifer Spielman, Barr C. Taylor, David Wolff, and Allan D. Wu, who were responsible for the design and construction of the AHTD device and organizing the trials in which the data used in this study was collected.

REFERENCES
