Expert Systems With Applications 46 (2016) 286–292

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Addressing voice recording replications for Parkinson's disease detection

Lizbeth Naranjo[a,*], Carlos J. Pérez[a], Yolanda Campos-Roca[b], Jacinto Martín[a]

[a] Department of Mathematics, University of Extremadura, Avda. de Elvas s/n, Badajoz 06006, Spain
[b] Department of Computer and Communication Technologies, University of Extremadura, Avda. de la Universidad s/n, Cáceres 10003, Spain

ARTICLE INFO

Keywords:
Bayesian binary regression
Gibbs sampling
Latent variables
Parkinson's disease
Speech recordings
Voice features

ABSTRACT

A clinical expert system has been developed for detection of Parkinson's Disease (PD). The system extracts features from voice recordings and considers an advanced statistical approach for pattern recognition. The significance of the work lies on the development and use of a novel subject-based Bayesian approach to account for the dependent nature of the data in a replicated measure-based design. The ideas under this approach are conceptually simple and easy-to-implement by using Gibbs sampling. Available information could be included in the model through the prior distribution. In order to assess the performance of the proposed system, a voice recording replication-based experiment has been specifically conducted to discriminate healthy people from people suffering PD. The experiment involved 80 subjects, half of them affected by PD. The proposed system is able to discriminate acceptably well healthy people from people with PD in spite that the experiment has a reduced number of subjects.

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease, affecting one in every 100 persons above the age of 65 years in Europe (De Rijk et al., 2000). According to the Parkinson's Disease Foundation, an estimated 7 to 10 million people worldwide are living with this medical condition. Depletion of dopaminergic nigrostriatal neurons gives rise to alterations in movement (tremor, rigidity, slow movements, unstable posture ...). Voice and speech, as dependent on movement of the articulators, are not spared. Non-dopaminergic changes can also affect language, cognition and mood, which can impact on communication (Miller, 2009).

The development of expert systems for medical diagnosis has received increasing attention in the literature for the last few decades (Singla, Grover, & Bhandar, 2014). These systems have the potential to optimize medical decisions, improve medical treatments, and reduce waiting lists and financial costs. In a typical disease diagnosis expert system, the core of the system is the knowledge base. Complex areas in medicine require extensive knowledge that may be extracted from clinical datasets (Fernandez-Millan, Medina-Merodio, Barchino, Martinez-Herraiz, & Gutierrez-Martinez, 2015; Halldorsson et al., 2015).

Traditional diagnosis of PD involves a physician taking a neurological history of the patient and performing an examination of a variety of motor skills. Since there is no definitive diagnostic test, the task is often difficult, particularly in the early stages when motor symptoms are not severe. Symptoms can be so subtle in these first stages that they go unnoticed, leaving the disease undiagnosed or misdiagnosed for extended periods of time. Clinical conditions leading to misdiagnosis or undiagnosis are one of the largest domains where medical expert systems receive increasing interest.

Voice recordings have been considered as a potential (non-invasive and low cost) biomarker to diagnose some voice-related diseases. Baghai-Ravary and Beet (2013) provided a current view of automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. Since the very early stages of PD, there can be subtle abnormalities in speech that might not be perceptible to listeners, but they could be evaluated in an objective way by performing acoustic analyses on recorded speech signals. Vocal impairment can be one of the earliest indicators of PD (Harel, Cannizzaro, & Snyder, 2004). Some authors have considered measures extracted from speech recordings to discriminate healthy people from those with PD (Little, McSharry, Hunter, Spielman, & Ramig, 2009; Sakar et al., 2013; Tsanas, Little, McSharry, Spielman, & Ramig, 2012).

The development of accurate remote systems considering features extracted from voice recordings can be very useful to help diagnose PD in its early stages. This idea may also help long-term undiagnosed and misdiagnosed patients. Besides, remote tracking of PD progression by considering voice recordings has also been considered in the scientific literature (Eskidere, Ertaç, & Hanilçi, 2012; Tsanas, Little, McSharry, & Ramig, 2010). The success of these systems would imply

* Corresponding author. Tel.: +34927257146; fax: +34924272911.
E-mail addresses: lizbeth@unex.es, lizbeth.naranjo.albarran@gmail.com (L. Naranjo), carper@unex.es (C.J. Pérez), ycampos@unex.es (Y. Campos-Roca), jrmartin@unex.es (J. Martín).

http://dx.doi.org/10.1016/j.eswa.2015.10.034
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

an improvement in patients' quality of life and a cost reduction for national health systems. Undoubtedly, there is a technological and scientific challenge to develop and disseminate expert systems for these tasks, so that they can be incorporated into protocols by neurological units.

Building a predictive model with minimal bias is intended to discriminate healthy people from people suffering PD, i.e., a model that maximizes the generalization of the predictions so as to perform well with new samples. In order to achieve this, a proper classification model must be considered. In this context, it has become usual to conduct experiments with replicated recordings. Little et al. (2009) presented one of the most used PD datasets consisting on 22 features extracted from 195 recordings of sustained /a/ phonations. These recordings belong to 32 people from both sexes, 24 of which were diagnosed with PD. Seven recordings were obtained from three subjects and six from the others, leading to an imbalanced design. This dataset is available online at UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Parkinsons). Hariharan, Polat, and Sindhu (2014) compare their proposals with the ones from fifteen previously published papers that use this dataset. Different overall accuracy rates were obtained depending on many factors, i.e., used features, reduction on features, classification methods, cross-validation schemes... A common point among all the used approaches is that they are based on independent sample schemes instead of on replicated measure-based frameworks. Note that each subject has six (or seven) replicated measures by feature which are not independent.

Independence-based classification methods should not be used when data have been obtained by replicating voice recordings from the same subjects. This fact artificially increases the sample size. Even more, this leads to a diffuse criterion to decide when a subject should be classified as suffering from PD, since it may happen (and it happens) that some voice recordings of the same person are classified as healthy and some others as disordered. For example, if the 195 22-dimensional vectors are used as training and testing datasets for a simple logistic regression, it is obtained that 4 out of 24 PD subjects and 4 out of 8 healthy subjects have different predictions in their own recordings. This means that 25% of the subjects (50% of the healthy and 16.7% of the PD) have incoherences among their own recording predictions. Note that, in this case, the global accuracy rate considering the recordings as independent would be 89.7% (72.9% for healthy and 95.2% for PD people).

Addressing dependent data as independent has become usual in this context. See, for example, Das (2010); Hariharan et al. (2014); Little et al. (2009); Tsanas et al. (2012); Von Orozco-Arroyave et al. (2013) and references therein. Sakar and Kursun (2010) noticed that traditional cross-validation methods divide recordings from the same individual in the training set and testing set, creating an artificial situation untypical of a real testing scenario. They defined an adapted cross-validation method named one-leave-individual-out. In this scheme, all the recordings of one individual are used for the testing set whereas all the recordings from the remaining individuals are used for the training set. This is performed for all the individuals and the accuracy rates are averaged. Although this is a positive step with respect to cross-validation, the underlying independence problem remains.

Silva, Dutra, Snasel, Platos, and El-Qawasmeh (2011) observed that this type of replicated data cannot be treated with traditional machine learning algorithms, since the data nature is dependent. They proposed to aggregate related data before learning by using some different functions as mean, minimum, maximum or a linear trend prediction. This leads to a clear criterion to discriminate between groups, avoiding the problem of defining which subject is healthy or not as happens with the other approaches. However, simplifying all the replicated measures from each subject into one single measure (for each feature) leads to an information loss. The within-subject variability is being removed by aggregating data. Therefore, the underlying within-subject dependence of the recordings must be properly modeled. Pérez, Naranjo, Martín, Campos-Roca, and EURASIP (2014) demonstrates the first classification approach for PD detection that takes into account the underlying within-subject dependence of the recordings by using the dataset provided in Little et al. (2009). The present work (performed by the same authors) is based on a new subject-based Bayesian probit approach that uses the idea of introducing latent variables to provide an efficient Gibbs sampling algorithm that overcome the computational issues.

Although, the classification approach is the main contribution in this paper, it is not the only one. We have built a system to extract features that are subsequently used in the classification approach to discriminate between healthy people and people with PD. Besides, a voice recording replication-based experiment has been specifically conducted to test the performance of this system by recording audio from both healthy people and people with PD.

The outline of this paper is as follows. The main information on participants, speech recordings, feature extraction and cross-validation methods is presented in Section 2. In Section 3, a subject-based Bayesian approach has been proposed. Section 4 presents the experimental results. In Section 5, a discussion on some specific points of the proposed system as well as its advantages and limitations is presented. Finally, Section 6 shows the conclusion.

2. Materials and methods

In this section, information on participants, speech recordings, feature extraction and cross-validation methods is presented. The approach is presented in the next section due to its extension.

2.1. Participants

A total of 80 subjects older than 50 years were involved in the study. 40 of them were healthy: 22 men (55%) and 18 women (45%), and 40 of them were affected by PD: 27 men (67.5%) and 13 women (32.5%). The mean (± standard deviation) age was 66.38 ± 8.38 for the control group and 69.58 ± 7.82 for the people with PD. PD patients presented at least two of the following symptoms: resting tremor, bradykinesia or rigidity.

The research protocol was approved by the Bioethical Committee from the University of Extremadura. All subjects signed an informed consent. The people with PD participating in this study were members of the Regional Association for Parkinson's Disease in Extremadura (Spain).

2.2. Speech recordings

The vocal task was the sustained phonation of /a/ vowel at comfortable pitch and loudness, as constant as possible. This phonation had to be kept for at least 5 seconds and on one breath. The task was repeated three times per individual, and all of them were considered as replications.

The speech data were recorded using a portable computer with an external sound card (TASCAM US322) and a headband microphone (AKG 520) featuring a cardiod pattern. The digital recording was performed at a sampling rate of 44.1 KHz and a resolution of 16 bits/sample by using Audacity software (release 2.0.5).

2.3. Feature extraction

The study is based on 44 acoustic features, which can be classified into five families: pitch local perturbation measures, amplitude local perturbation measures, noise features, spectral envelope measures and nonlinear ones.

Four pitch local perturbation measures were obtained: jitter relative (expressed in percentage), jitter absolute, jitter RAP (Relative

Average Perturbation) and jitter PPQ (Pitch Perturbation Quotient) (Baken & Orlikoff, 2000). These features were extracted by using a waveform matching algorithm consisting in two steps: (1) Calculation of the rough fundamental period length by using the normalized auto-correlation function over 80 ms frames and, (2) Second application of the auto-correlation function on segments with a length equal to twice the mean value of the fundamental periods roughly estimated before, after removing gross pitch errors (halving and doubling) and unvoiced frames. Waveform-matching algorithms have been proved to outperform peak-picking methods for jitter estimation (Titze & Liang, 1993).

The five amplitude perturbation measures are: shimmer local, shimmer dB, APQ3 (3-point Amplitude Perturbation Quotient), APQ5 (5-point Amplitude Perturbation Quotient) and APQ11 (11-point Amplitude Perturbation Quotient) (Baken & Orlikoff, 2000).

Harmonic-to-noise ratio (HNR) is a measure of the relative level of noise present in speech. There are many variants of HNR (based on time-domain or frequency-domain approaches). In this work we have used 5 different HNR features, corresponding to different frequency bandwidths: HNR05 (0-500 Hz), HNR15 (0-1500 Hz), HNR25 (0-2500 Hz), HNR35 (0-3500 Hz) and HNR38 (0-3800 Hz). Individual HNR values for each frame are extracted by using the VoiceSauce toolbox (Shue, Keating, Vicenik, & Yu, 2010). These HNR measures are calculated using a cepstrum-based technique (Krom, 1993). The final HNR features are calculated as average values of all voiced frames. To the authors' knowledge, the use of several HNR measures based on different bandwidths has not been reported yet for PD detection.

Glottal-to-Noise Excitation Ratio (GNE) attempts to quantify the amount of voice excitation by vocal-fold oscillations versus excitation by turbulent noise (Michaelis, Gramss, & Strube, 1997). It is also included in the family of noise features. An advantage of GNE feature is that its calculation is not based on a previous estimation of the fundamental frequency, which is always a critical step in the presence of pathology.

Mel Frequency Cepstral Coefficients (MFCCs) are related to the speech spectral envelope, which depends on articulator position. PD is known to affect also articulation, therefore this type of coefficients are promising features to characterize PD (Tsanas, Little, McSharry, & Ramig, 2011). By the use of MFCCs, it should be possible to detect slight misplacement of the articulators. Delta MFCC features are time derivatives of MFCCs, so they can be used to detect subtle changes in the articulator positions (due to tremor) when performing a sustained fonation. In this work, both types of coefficients have been used to characterize PD. The length of the feature vector has been chosen to be 26 (13 MFCCs plus 13 Delta coefficients).

The existence of non-linear phenomena in the production process of the speech signal has been theoretically and experimentally established. Nonlinear features are of particular relevance to clinical practice, because severe dysphonic pathological voices are precisely the ones that are most likely to present highly nonlinear and random phenomena, whereas healthy voices are closer to the linear source-filter model. The following nonlinear features have been considered in this work: RPDE (Recurrence Period Density Entropy) (Little, McSharry, Roberts, Costello, & Moroz, 2007), DFA (Detrended Fluctuation Analysis) (Little et al., 2007) and PPE (Pitch Period Entropy) (Little et al., 2009).

After extraction, a matrix with 240 rows (80 subjects × 3 replications) and 44 columns (one for every voice feature) is obtained.

2.4. Classification model and validation

A subject-based Bayesian classification approach will be proposed in Section 3. Since each subject has some dependent measurements for each voice variable, the approach must be able to address recording replications in a proper way. The approach will be presented in a single section because of the extension.

Section 4 shows the experimental results. Firstly, all the individuals will be used as a training set and as a testing set. This will provide an idea on the model performance. However, the closest-to-reality approach is to use all the individuals as a training set and use the model parameters to predict the health status of future subjects. Since there are only 80 subjects whose health status is well known and no new individuals are available, cross-validation will be used in order to assess the model generalization performance (Webb, 2002). In this case, the experimental unit is the subject and not the recording. When applying cross-validation with the proposed approach the success rate is based on subjects containing their three recordings replications. In this way, the prediction will be presented for each subject and not for each recording. Specifically, the dataset is randomly split into a training subset composed by 75% of the control subjects and 75% of the people with PD. The remaining individuals constitute the testing subset. The model parameters are determined using the training subset, and errors are computed using the testing subset. This is performed 100 times and the results are then averaged.

3. A latent variable-based approach

In this section, a generalized linear approach with latent variables is introduced to address replicated measurements in a general classification context. Specifically, hierarchical binary regression models are considered from a Bayesian viewpoint. In Bayesian methodology, the initial knowledge about the parameters (prior distribution) is combined with the model considering the observed data (likelihood) to provide the posterior distribution. The posterior distribution contains all the information about the model parameters. This methodology allows that the initial information from historical data or experts can be included in the model through the prior distribution. This can be a great advantage when information different from the current data is available. The proposed approach allows the inclusion of initial information for both the model parameters and the covariates. If no information is available, flat distributions can be used instead.

3.1. Model description

Suppose that n independent binary random variables $Y_1, \ldots, Y_n$ are observed, where Y_i is Bernoulli distributed with success probability $\mathrm{P}(Y_i = 1) = p_i,\ i = 1, \ldots, n$. The probabilities p_i are related to two sets of covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, where $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iK})^t$ is a $K \times J$ matrix of a set of K covariates which have been measured with J replicates, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH})^t$ is a H vector of a set of H covariates which are exactly known. Suppose that $\boldsymbol{x}_{ij} = (x_{i1j}, \ldots, x_{iKj})$ is the jth replication of the unknown covariates vector $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iK})$ and assume that they have a linear relationship (additive measurement error model, see e.g. Buonaccorsi (2010)), i.e., instead of the covariates $\boldsymbol{w}$, it is observed their replicates $\boldsymbol{x}$. By this way the $\boldsymbol{x}_{ij}$'s are the surrogates of $\boldsymbol{w}_i$'s. The parameters p_i are related to $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_i$ through the following hierarchical model

$$\begin{aligned} Y_i &\sim \mathrm{Bernoulli}(p_i), \\ \Psi^{-1}(p_i) &= \boldsymbol{w}_i^t \boldsymbol{\beta}_x + \boldsymbol{z}_i^t \boldsymbol{\beta}_z, \\ \boldsymbol{x}_{ij} &= \boldsymbol{w}_i + \boldsymbol{\varepsilon}_{ij}, \\ \boldsymbol{\varepsilon}_{ij} &\sim \mathrm{Normal}_K(\boldsymbol{0}, \boldsymbol{G}), \end{aligned} \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^t, \boldsymbol{\beta}_z^t)^t$ is a $(K + H)$ vector of unknown parameters, $\Psi^{-1}(\,\cdot\,)$ is a known nonnegative and nondecreasing function ranging between 0 and 1 (called the link function), and $\boldsymbol{G}$ is a $K \times K$ matrix of variances and covariances (the replicates between covariates are not independent). The error vector $\boldsymbol{\varepsilon}_{ij}$ is independent of $\boldsymbol{w}_i$, implying that $\boldsymbol{x}_{ij}$ is a surrogate of $\boldsymbol{w}_i$. Usually $\Psi(\,\cdot\,)$ is the cumulative distribution function of the normal or logistic distribution (probit and logit link functions, respectively).

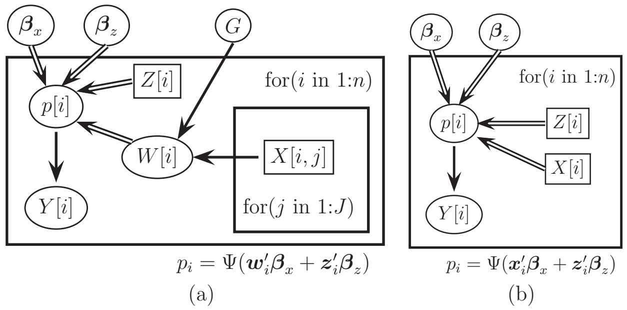
β_x
β_z
G
$Z[i]$
for(i in 1:n)
$p[i]$
$W[i]$
$X[i,j]$
for(j in 1:J)
$Y[i]$
$p_i = \Psi(\boldsymbol{w}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z)$
(a)
β_x
β_z
for(i in 1:n)
$p[i]$
$Z[i]$
$X[i]$
$Y[i]$
$p_i = \Psi(\boldsymbol{x}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z)$
(b)

Fig. 1. Flowcharts for the proposed approach (a) and for the standard approach (b).

The following step is to define the prior distributions. A usual approach for the regression models assumes a multivariate normal distribution for the regression parameters, $\boldsymbol{\beta} \sim \text{Normal}_{K+H}(\boldsymbol{b}, \boldsymbol{B})$, and it is assumed conjugate prior distributions for the variance and covariance parameters, $\boldsymbol{G} \sim \text{InvWishart}_K(\boldsymbol{V}, \nu)$, where $\boldsymbol{b}$**,** $\boldsymbol{B}$, ν and $\boldsymbol{V}$ are fixed. Also, suppose that the latent variables are distributed from $\boldsymbol{w}_i \sim \text{Normal}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are also fixed values.

The likelihood function considering the observed and latent variables is

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) = f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}) f(\boldsymbol{x}|\boldsymbol{w}, \mathbf{G}) f(\boldsymbol{w}),$$

then, the joint posterior density is

$$\pi(\boldsymbol{\beta}, \mathbf{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) \propto \mathcal{L}(\boldsymbol{\beta}, \mathbf{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) \pi(\boldsymbol{\beta}) \pi(\mathbf{G}).$$

This approach uses the relationship between the covariates and the latent variables jointly with the prior distributions to achieve posterior estimations of the latent variables. The approach considers the latent variables as missing data and provides imputations from the distribution conditioned on the observed variables and the parameters. The flowcharts displayed in Fig. 1 represent both the proposed binary regression model and the standard model.

In the standard binary regression model proposed by Albert and Chib (1993), the variables $\mathbf{x}_i$'s are exactly known, i.e., the voice measures should be exactly known without any replication. However, in the proposed model, the $\boldsymbol{x}_{ij}$'s are replications of the voice measures, and they are the surrogates of $\boldsymbol{w}_i$'s. These $\boldsymbol{w}_i$'s are estimated and, at the same time, used in the proposed model. By tackling the problem in this way, replicated measurement can be considered, whereas the standard method cannot. This allows to classify people with PD at the same time that within-subject variability is taken into account.

3.2. Exploring the posterior distribution

The joint posterior density is not tractable for computing, so it must be estimated. MCMC methods are computing tools that can be used in this context (Gilks, Richarson, & Spiegelhalter, 1995). WinBUGS software has been widely used to implement MCMC simulations (Ntzoufras, 2011). This computing environment has the advantage that it is very easy to program and the code is easily adapted for different link functions. For example, a particular case of the generalized linear approach has been presented by Pérez et al. (2014) considering this computing environment for the logistic case. However, by using WinBUGS the user is not able to exactly know how the generation process is being performed. An specific algorithm has been developed and implemented in R language (https://cran.r-project.org/).

In this section, we propose to use the probit link function, i.e. the inverse of the normal cumulative distribution function. By considering the idea of introducing latent variables to produce a data augmentation framework, the use of the probit link function leads to the development of an efficient Gibbs sampling algorithm with easy-to-generate full conditional distributions (necessary to implement the iterative process). For example, if the logistic case is considered, then the full conditional distributions are not easy to generate from and Metropolis-Hastings methods must be used. This makes the generation process more difficult.

Latent variables are introduced based on the proposal of Albert and Chib (1993), i.e., n independent latent variables $u_1, \ldots, u_n$ are considered, where u_i is distributed as

$$u_i \sim \text{Normal}(\boldsymbol{w}_i^t \boldsymbol{\beta}_x + \boldsymbol{z}_i^t \boldsymbol{\beta}_z, 1),$$

and it is defined

$$Y_i = \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{if } u_i \leq 0 \end{cases},$$

and therefore

$$\text{P}(Y_i = y_i) = \int_{-\infty}^{\infty} (2\pi)^{-1/2} \exp\left\{ -\tfrac{1}{2}(u_i - \boldsymbol{w}_i^t \boldsymbol{\beta}_x - \boldsymbol{z}_i^t \boldsymbol{\beta}_z)^2 \right\} \times \left\{ I[y_i = 1] I[u_i > 0] + I[y_i = 0] I[u_i \leq 0] \right\} \text{d}u_i.$$

Then, the likelihood function including the latent variables is

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{w}) = f(\boldsymbol{y}|\boldsymbol{u}) f(\boldsymbol{u}|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}) f(\boldsymbol{x}|\boldsymbol{w}, \mathbf{G}) f(\boldsymbol{w}),$$

so the posterior distribution is

$$\pi(\boldsymbol{\beta}, \mathbf{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{w}) \propto \mathcal{L}(\boldsymbol{\beta}, \mathbf{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{w}) \pi(\boldsymbol{\beta}) \pi(\mathbf{G}).$$

Then, the full conditional distributions are given by

$$u_i |\, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}, \mathbf{G} \sim \begin{cases} \text{Normal}(\boldsymbol{w}_i^t \boldsymbol{\beta}_x + \boldsymbol{z}_i^t \boldsymbol{\beta}_z, 1) I[u_i > 0] & \text{if } y_i = 1 \\ \text{Normal}(\boldsymbol{w}_i^t \boldsymbol{\beta}_x + \boldsymbol{z}_i^t \boldsymbol{\beta}_z, 1) I[u_i \leq 0] & \text{if } y_i = 0 \end{cases}, \tag{2}$$

$$\boldsymbol{w}_i |\, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{\beta}, \mathbf{G} \sim \text{Normal}_K(\boldsymbol{m}_i, \boldsymbol{M}), \tag{3}$$

$$\boldsymbol{\beta} |\, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{w}, \mathbf{G} \sim \text{Normal}_{K+H}(\boldsymbol{b}^*, \boldsymbol{B}^*), \tag{4}$$

$$\mathbf{G} |\, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{\beta} \sim \text{InvWishart}(\boldsymbol{V}^*, nJ + \nu), \tag{5}$$

where

$$\begin{aligned} \boldsymbol{m}_i &= \boldsymbol{M}\left(\boldsymbol{\beta}_x (u_i - \boldsymbol{z}_i^t \boldsymbol{\beta}_z) + \sum_{j=1}^{J} \mathbf{G}^{-1} \boldsymbol{x}_{ij} + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right), \\ \boldsymbol{M} &= \left(\boldsymbol{\beta}_x \boldsymbol{\beta}_x^t + J\mathbf{G}^{-1} + \boldsymbol{\Sigma}^{-1} \right)^{-1}, \\ \boldsymbol{b}^* &= \boldsymbol{B}^* \left[(\boldsymbol{w}, \boldsymbol{z})^t \boldsymbol{u} + \boldsymbol{B}^{-1} \boldsymbol{b} \right], \\ \boldsymbol{B}^* &= \left[(\boldsymbol{w}, \boldsymbol{z})^t (\boldsymbol{w}, \boldsymbol{z}) + \boldsymbol{B}^{-1} \right]^{-1}, \\ \boldsymbol{V}^* &= \sum_{i=1}^{n} \sum_{j=1}^{J} (\boldsymbol{x}_{ij} - \boldsymbol{w}_i)(\boldsymbol{x}_{ij} - \boldsymbol{w}_i)^t + \boldsymbol{V}. \end{aligned}$$

The final Gibbs sampling-based algorithm consists of choosing initial values $w^{(0)}$, $\beta^{(0)}$ and $G^{(0)}$, and iteratively sampling $u^{(l)}$, $w^{(l)}$, $\beta^{(l)}$ and $G^{(l)}$ from the full conditional distributions (2)–(5), respectively.

3.3. Estimating the predictive probabilities

Probability predictions for future observations y^* are based on the predictive distributions. When y has not been observed yet, predictions are based on the marginal likelihood $f(y^*|\theta)$ and the prior distribution $f(\theta)$, obtaining the prior predictive distribution

$$f(\boldsymbol{y}^*) = \int f(\boldsymbol{y}^*|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \text{d}\boldsymbol{\theta},$$

where θ is the set of all the parameters in the model. After having observed data y, the predictions of future observations y^* are based on the marginal likelihood of the future observations $f(y^*|\theta)$ and the posterior distribution $f(\theta|y)$ through the posterior predictive distribution

$$f(\boldsymbol{y}^{*}|\boldsymbol{y}) = \int f(\boldsymbol{y^{*}}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}.$$

In order to obtain predictions in the proposed model, latent variables w_i are generated from the multivariate normal distribution $\text{Normal}_K(m_i, M)$, where

$$\boldsymbol{m}_i = \boldsymbol{M}\left(\sum_{j=1}^{J} \boldsymbol{G}^{-1}\boldsymbol{x}_{ij} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right),$$

$$\boldsymbol{M} = \left(J\boldsymbol{G}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}.$$

Then, it is computed $p_i = \Psi(\boldsymbol{w}_i^t \boldsymbol{\beta}_x + \boldsymbol{z}_i^t \boldsymbol{\beta}_z)$, and $y_i \sim \text{Bernoulli}(p_i)$. This is the way how samples from the posterior predictive distribution of p_i and y_i are obtained. Note that G, β_x and β_z are the posterior samples generated in the Gibbs sampling.

4. Experimental results

In this section, the acoustic variables were individually normalized to have mean 0 and standard deviation 1. The response variable Y takes values $y = 1$ for people with PD and $y = 0$ for healthy subjects. The following hyperparameters were used:

- μ represents the sample mean vector of the normalized covariates, so all its elements are zeros.
- Σ is the sample correlation matrix of the normalized variables, so it is a matrix composed of a diagonal of ones.
- The prior distribution of $\boldsymbol{\beta}_{(K+H)}$ is a normal distribution with mean $\boldsymbol{b} = (0, \dots, 0)^t$ and covariance matrix equal to $\boldsymbol{B} = \text{diag}_{K+H}(100)$.
- The hyperparameters of G are $\nu_0 = K$ and $V_0 = \text{diag}_K(1)$.

Firstly, the proposed method is applied with all the subjects as the training set and all the subjects as the test set. The accuracy rate obtained from the predictive distribution is 0.850. Recall (TP/(TP+FN)), specificity (TN/(TN+FP)), and precision (TP/(TP+FP)) are 0.800, 0.900 and 0.889, respectively. Note that these results have been obtained with a model that considers replications, and the predictions have been obtained for 80 subjects (with their three replications included in the model). If a standard probit model is applied to the 240 recordings (considering all recordings as independent), it is obtained an accuracy rate of 0.871 (recall=0.842, specificity=0.900, and precision=0.894). In spite that the accuracy rate is greater than the one obtained with the proposed approach, it has been obtained that 22 subjects (27.5%) have incoherences in the predictions of their own three recordings, i.e. some recordings of the same individual are predicting healthy and some are predicting PD (independently of whether the prediction is correct or not). Specifically, this happens for 22.5% (9 out of 40) of the healthy subjects and 32.5% (13 out of 40) of the people with PD. This shows that artificially increasing the sample size may provide better accuracy rates, but leads to an incoherent criterion for the decision making. The proposed approach avoids this fact at the same time that provides a statistically rigorous method with good accuracy rates.

In order to validate the results, a stratified cross-validation framework is considered. Specifically, the dataset is randomly split into a training subset composed by 75% of the control subjects and 75% of the people with PD. The remaining individuals constitute the testing subset. The model parameters are determined using the training subset, and errors are computed using the testing subset. This is achieved 100 times and the results are then averaged. Table 1 presents the results.

Note that the accuracy rate is 0.752 and the standard deviation is 0.086. This model has a moderate predictive capacity in absolute terms, but in this context, the approach is providing a good accuracy rate, since the approach is based on subjects and no artificial increment of the sample size has been considered. The within-subject variability has been considered in this approach.

Table 1
Accuracy rate and other indicators.

All	Mean	SD
Accuracy rate	0.752	0.086
Recall	0.718	0.132
Specificity	0.786	0.135
Precision	0.785	0.118

Table 2
Accuracy rates and other indicators divided by sex.

Men	Mean	SD
Accuracy rate	0.706	0.110
Recall	0.667	0.157
Specificity	0.760	0.182
Precision	0.811	0.130
Women	Mean	SD
Accuracy rate	0.876	0.122
Recall	0.873	0.205
Specificity	0.878	0.161
Precision	0.853	0.185

The same cross-validation scheme is considered for men and women separately. Note that there are 49 men and 31 women. Then, the training sets are composed of 36 men and 24 women, respectively, whereas the testing sets are composed of only 13 men and 7 women, respectively. Table 2 shows the results.

Note that the accuracy rate for men is lower than the one obtained when all the subjects were considered independently of the sex, whereas the accuracy rate for women is much larger. These results are remarkable due to the reduced available sample size. This reduced sample size has the effect of providing large standard deviations.

As a summary, the proposed approach is able to discriminate relatively well healthy subjects from people with PD in an experiment with a reduced sample size. This discrimination capacity is even greater for women than for men.

5. Discussion

Many experimental data are collected in replicated measure-based statistical designs. Here the term 'replications' refers to the collection of features extracted from voice recordings belonging to the same subject. Since, in this context, features are extracted from multiple voice recordings from the same subject, in principle, the features should be identical. The imperfections in technology and the own biological variability result in non-identical replicated features that are more similar to one another than features from different subjects.

Traditional statistical approaches based on repeated measurements do not properly fit this experimental design because they refer to situations in which the response of each experimental unit or subject is observed in multiple occasions or under multiple conditions. Then, the concept of replication considered here does not match the classical concept of statistical repeated measurements (Stroup, 2013).

One key point in the development of the proposed clinical expert system for PD detection is the classification method. Failure to account for replications may result in a classification approach producing overoptimistic estimated accuracy rates, making subsequent classifications unreliable. Ignoring the dependent nature of the observations has become usual for PD detection (see, for example, Hariharan et al. (2014); Little et al. (2009); Tsanas et al. (2012); Von Orozco-Arroyave et al. (2013) and references therein). Conventional machine learning methods are not appropriate since observations are not independent. Some authors have aggregated replicated data before learning by using some different functions for PD detection (see, e.g., Sakar et al. (2013); Silva et al. (2011)). However, a loss of information is incurred because the within-subject variability is fully removed.

Alternative modeling approaches must be used to avoid the mentioned problems and provide a rigorous classification methodology matching the right experimental design. Pérez et al. (2014) demonstrates the first classification approach for PD detection that takes into account the underlying within-subject variability by using the dataset provided in Little et al. (2009). This classification approach was based on a logistic regression model and was implemented in WinBUGS (Ntzoufras, 2011). This software has the disadvantage of using *black-boxes* in the generation process, what does not allow the user to know how this process is being performed. However, this new classification approach proposed here (developed by the same authors) is based on the probit link function and considers the introduction of latent variables as in Albert and Chib (1993) to derive an efficient Gibbs sampling algorithm. The full conditional distributions necessary to implement the iterative process are explicitly provided and they are ready for use. An MCMC convergence assessment is necessary in order to validate the results. If the logistic case is considered, then the generation process involves Metropolis–Hastings methods, what makes the generation process more difficult. Note that the acceptance rate in Gibbs sampling is always 100%, whereas the acceptance rate for Metropolis–Hastings depends on a proposal distribution. This may make the algorithm more inefficient.

Lee, Kuo, Whitmore, and Sklar (2000) showed the importance of replications in a different context, specifically, in microarray gene expression studies. Karpievitch, Hill, Leclerc, Dabney, and Almeida (2009) proposed a classifier (RF++) capable of analyzing cluster-correlated data. It was developed as an implementation of the random forest algorithm, as described by Breiman (2001), with additional functionality specific to the structure of cluster-correlated data. The approach is implemented in the following steps: (i) RF++ grows each tree on a bootstrap sample at subject level rather than at replication level of the training data. Individual trees are unpruned decision trees grown using Gini impurity score (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). A particular subject is chosen at random from the pool of all available subjects and all of its replications are allocated to an in-bag dataset. Approximately 63% of the individual samples are in-bag and the remainder are held out in order to compute a runtime error estimate on the out-of-bag samples; (ii) A means for computing subject-level classification is provided. Specifically, subject replications at the sample-level are first classified and then a majority vote across subject replications is performed in order to compute a classification; (iii) RF++ provides an error rate based on out-of-bag data. When all subject replications belong to the same class, unbiased running out-of-bag subject-level error estimate is computed.

A first difference between this approach and ours is that the former belongs to the nonlinear classifier class, whereas the latter belongs to the linear classifier class. The main advantage of the proposed approach over RF++ is that RF++ is based on votes so there may be incoherences among the own classification of the replications for each individual. Karpievitch et al. (2009) commented that occasional misclassifications of the replications on one concrete individual (e.g., misclassifications of one or two replications out of a collection of replications) generally have little effect on the final forest subject-level error rate. However, there may be important effects when the percentage of misclassifications is not so low. The problem of incoherences among the replications of each subject has been treated here for both the data in Little et al. (2009) and the data obtained from this experiment, leading to the conclusion that vote-based methods may provide high percentages of incoherences among the own replications of each individual. Finally, it is remarkable that RF++ is a multi-class classifier, whereas the proposed classifier has been designed for two-class problems. Although the proposed approach can be easily extended to multi-class problems, no testing has been performed yet, since the main application has been PD discrimination between healthy subjects and people with PD.

Table 3
Strengths and weaknesses of the proposed classification approach.

Strengths
- Properly fits the experimental design. - Takes into account the within-subject variability. - Allows to include prior information. - Derives an efficient Gibbs sampling-based algorithm. - Classifies individuals and not replications.
Weaknesses
- More variability than aggregating replications. - More computational cost than aggregating replications. - An MCMC convergence assessment must be performed. - Implemented and tested for only two-class problems.

Table 3 presents a summary of the strengths and weaknesses of the proposed classification approach.

The clinical expert system developed also contains a feature extraction procedure. The success of the classification also highly depends on the good discriminatory properties of the voice features. We have applied a number of features, both linear and nonlinear ones. Some of them have been used for the first time in PD detection as the HNR based on different bandwidths. Parkinsonian voices often show an increase of turbulence because of incomplete vocal fold closure leading to an abnormally high degree of hoarseness. Thus, the research community has addressed the issue of automatic detection of PD by including the HNR parameter into the set of discriminating features (Little et al., 2009). However, this feature considers the ratio between harmonic and noise energy over the full bandwidth. It does not consider the role of specific frequency bands. Subband-based HNR parameters have been recently proposed as a measure for the assessment of creaky voice (Keating & Garellek, 2015). However, to the authors' knowledge, they have not been tested for discrimination of PD yet.

In order to assess the performance of the proposed system, a recording replication-based experiment has been specifically conducted to discriminate healthy people from people suffering PD. The experiment involved 80 subjects (40 affected by PD and the other ones were healthy). Conducting this kind of experiment is difficult due to the need to recruit people suffering PD who voluntarily perform the proposed vocal tasks. We counted with the collaboration of the Regional Association for Parkinson's Disease in Extremadura (Spain). Although the number of subjects is moderate, the proposed system is able to discriminate acceptably well healthy subjects from people with PD. The accuracy rate is 85.0% when considering all the subjects as training and testing sets. This percentage reduces to 75.2% when cross-validation is considered. This discrimination is even better for women (87.6%) than for men (70.6%), what confirms the need to apply the system separately for women and men. Hertrich and Ackermann (1995) point out that PD seems to have a differential impact on phonation in men and women and that these gender-specific vocal dysfunctions may be explained by the different laryngeal size. Although measures of dysprosody are not considered in this investigation, gender-specific impact of PD on speech prosody has also been reported (Skodda, Visser, & Schleger, 2011).

6. Conclusion

An expert system that discriminates people with PD from healthy controls based on acoustic features extracted from voice recordings has been developed. Through a novel subject-based Bayesian classification approach the system is able to account for the dependent nature of the data in a replicated measure-based design. The proposed approach has been implemented and tested for two-classes classification problems, but it can be easily extended to the multi-class problem. It has been developed to be included in this system related to PD detection, but it can be used in similar experiments for different purposes.

Although the main contribution is the classification approach, the system also extracts voice features according to different algorithms. These algorithms provide the data for the classification approach. The improvement or definition of new extraction procedures is a non-ending task, since the more discriminative the features are, the better accuracy rates will be obtained. The obtained results show that the whole system provides good results, however they have a margin for improvement by developing new feature extraction algorithms.

A future research in expert systems for PD detection will consist in the development of an intelligent telediagnosis system prototype based on mobile terminal devices. Smartphones are emerging as a low-cost and feasible technology for the rapid growth of m-health (mobile health) applications. This fact can be used to extend the proposed system to a real-time one. A goal for the immediate future is to increase the database by recording new subjects considering adverse recordings conditions, such as environmental noise or different device-to-mouth distances. This process will be performed by opening the system to other associations of Parkinson.

Acknowledgments

Thanks to the anonymous participants and to Carmen Bravo and Rosa María Muñoz for carrying out the voice recordings and providing information from the people with PD. We also thank two referees for comments and suggestions which have highly improved both the readability and the content of this paper. We are grateful to the *Asociación Regional de Parkinson de Extremadura* and *Confederación Española de Personas con Discapacidad Física y Orgánica* for providing support in the experiment development.

This research has been supported by *Ministerio de Economía y Competitividad*, Spain (Projects MTM2011-28983-C03-02 and MTM2014-56949-C3-3-R), *Gobierno de Extremadura*, Spain (Projects GR15052 and GR15106), and *European Union* (European Regional Development Funds).

References

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*, 669–679.

Baghai-Ravary, L., & Beet, S. W. (2013). *Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. Springer briefs in electrical and computer engineering - speech tecnology*. New York: Springer.

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

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods and applications*. Boca Raton, Florida: Chapman and Hall/CRC.

Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson's disease. *Expert Systems with Applications, 37*, 1568–1572.

De Rijk, M. C., Launer, L. J., Berger, K., Breteler, M. M., Dartigues, J. F., Baldereschi, M., … Hofman, A. (2000). Prevalence of Parkinson's disease in Europe: a collaborative study of population-based cohorts. *Neurology, 54*, S21–S23.

Eskidere, O., Ertaç, F., & Hanilçi, C. (2012). A comparison of regression methods for remote tracking of Parkinson's disease progression. *Expert Systems with Applications, 39*, 5523–5528.

Fernandez-Millan, R., Medina-Merodio, J. A., Barchino, R., Martinez-Herraiz, J. J., & Gutierrez-Martinez, J. M. (2015). A laboratory test expert system for clinical diagnosis support in primary health care. *Applied Science, 5*, 222–240.

Gilks, W. R., Richarson, S., & Spiegelhalter, D. J. (1995). *Markov chain Monte Carlo in practice*. London: Chapman & Hall/CRC.

Halldorsson, B. V., Bjornsson, A. H., Gudmundsson, H. T., Birgisson, E. O., Ludviksson, B. R., & Gudbjornsson, B. (2015). A clinical decision support system for the diagnosis, fracture risks and treatment of osteoporosis. *Computational and Mathematical Methods in Medicine*, 1–7.

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 and Cognition, 56*, 24–29.

Hariharan, M., Polat, K., & Sindhu, R. (2014). A new hybrid intelligent system for accurate detection of Parkinson's disease. *Computer Methods and Programs in Biomedicine, 113*, 904–913.

Hertrich, I., & Ackermann, H. (1995). Gender-specific vocal dysfunctions in parkinson's disease: electroglottographic and acoustic analyses. *Annals of Otology, Rhinology and Laryngology, 104*, 197–202.

Karpievitch, Y. V., Hill, E. G., Leclerc, A. P., Dabney, A. R., & Almeida, J. S. (2009). An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. *PLoS One, 4*, e7087(1–10).

Keating, P., & Garellek, M. (2015). Acoustic analysis of creaky voice. In *Annual meeting of the linguistic society of America*.

Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language and Hearing Research, 36*, 254–266.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Science U.S.A., 97*, 9834–9839.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering, 56*, 1015–1022.

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A. E., & Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine, 6*, 1–19.

Michaelis, D., Gramss, T., & Strube, H. W. (1997). Glottal-to-noise excitation ratio - a new measure for describing pathological voices. *Acta Acustica united with Acustica, 83*. 700–706(7)

Miller, N. (2009). Communication changes in Parkinson's disease. *Revista de Logopedia, Foniatría y Audiología, 29*, 37–46.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS. Wiley series in computational statistics*. New Jersey: Wiley.

Pérez, C. J., Naranjo, L., Martín, J., & Campos-Roca, Y. (2014). A latent variable-based Bayesian regression to address recording replication in Parkinson's disease. In EURASIP (Ed.), *Proceedings of the 22nd European signal processing conference (EUSIPCO-2014)* (pp. 1447–1451). Lisbon, Portugal: IEEE.

Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., … Kursun, O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics, 17*, 828–834.

Sakar, C. O., & Kursun, O. (2010). Telediagnosis of Parkinson's disease using measurements of dysphonia. *Journal of Medical Systems, 34*, 591–599.

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2010). VoiceSauce: A program for voice analysis. *Energy, 1*. H1–A1

Silva, T., Dutra, I., Snasel, V., & Platos, J. (2011). T-SPPA trended statistical preprocessing algorithm. In E. El-Qawasmeh (Ed.), *The international conference on digital information processing and communications: I* (pp. 118–131). Springer-Verlag.

Singla, J., Grover, D., & Bhandar, A. (2014). Medical expert systems for diagnosis of various diseases. *International Journal of Computer Applications, 93*, 36–43.

Skodda, S., Visser, W., & Schleger, U. (2011). Gender-related patterns of dysprodosy in Parkinson disease and correlation between speech variables and motor symptoms. *Journal of Voice, 25*, 76–82.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics, 8*.

Stroup, W. W. (2013). *Generalized lineal mixed models: modern concepts, methods and applications*. Boca Raton, Florida: Chapman and Hall/CRC.

Titze, I. R., & Liang, H. (1993). Comparison of F0 extraction methods for highprecision voice perturbation measurements. *Journal of Speech, Language and Hearing Research, 36*, 1120–1133.

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *IEEE Transactions Biomedical Engineering, 57*, 884–893.

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *The Royal Society Interface, 8*, 842–855.

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Transactions on Biomedical Engineering, 59*, 1264–1271.

Von Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., & Nöth, E. (2013). Analysis of speech from people with Parkinson's disease through nonlinear dynamics. In T. Drugman, & T. Dutoit (Eds.), *Advances in nonlinear speech processing*. In *Lecture Notes in Artificial Intelligence. Subseries of Lecture Notes in Computer Science: LNAI 7911* (pp. 112–119). Springer-Verlag.

Webb, A. (2002). *Statistical pattern recognition*. Chichester: John Wiley and Sons.