Estimating the Severity of Parkinson’s Disease from Speech
Using Linear Regression and Database Partitioning

Dávid Sztahó, Gábor Kiss, Klára Vicsi
Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
sztaho@tmit.bme.hu, kiss.gabor@tmit.bme.hu, vicsi@tmit.bme.hu

Abstract
Parkinson’s disease (PD) is one of the most common neurodegenerative disorders. PD is referred as idiopathic, that is, as having no known cause; its main symptoms are tremor, rigidity and general loss of muscle control. Research shows that speech may be a useful indicator for discriminating patients with PD from healthy controls. The paper describes our contribution to the INTERSPEECH 2015 Special Session “Computational Paralinguistics Challenge (ComParE): Parkinson’s Condition Sub-Challenge”. The main goal of the challenge is to perform automatic classification (regression) on speech produced by patients with Parkinson’s disease. The paper presents our method of linear regression models on a set of extracted acoustic features from the middle of vowels in words, sentences and continuous speech, and the partitioning of the speech samples according to their total length into parts with long, medium and short duration.

Index Terms: Parkinson’s disease, speech analysis, machine learning, automatic regression

1. Introduction
Parkinson’s disease (PD) is one of the most common neurodegenerative disorders with an incidence rate of approximately 20/100,000 [2]. In Hungary the estimated number of people suffering from Parkinson’s disease is in the range of about 16-20,000. The main cause of Parkinson’s disease is the damage or death of neurons that produce dopamine in the area of the brain called substantia nigra. Dopamine is a neurotransmitter that participates in the smooth, flicker-free, fluent regulation of skeletal muscles (muscles that execute voluntary movements). Although the number of neurons that produces dopamine decreases naturally with age, if at least half of the initial number remains, the disease will not develop and the potential shaking of the hands will be of a low degree. However, the above-mentioned neurons in the substantia nigra are sensitive to adverse chemical influences, thus providing a basis for PD to develop. The main risk factors of PD are age, genetics and medical history, but the most important of these is age. This fact is rather important because in a continuously aging society the number of patients may also increase and therefore it may be important to be able to detect the disease at an early stage and/or by using a simple and fast procedure.

Main symptoms of PD include tremor, rigidity and general loss of muscle control, as well as cognitive impairment. Additional complications can arise, such as depression, dementia, hard swallowing, constipation and sleep disorders. During diagnosis the therapist evaluates the patient’s mental state, muscle strength, movement coordination, reflexes and sensory skills. However, research [3][4] shows that speech may be a useful indicator for discriminating patients with PD from healthy controls. Clinical evidence suggests that most PD patients have some form of speech disorder [4], and therefore speech can be an early sign of PD [6]. Informal descriptions of symptoms that can characterize the speech of PD patients include reduced loudness, increased vocal tremor and breathiness. The two vocal impairments linked to PD are dysphonia (inability to produce normal vocal sounds) and dysarthria (difficulty in pronouncing words) [7].

The intention of diagnosing PD from speech at an early stage has inspired researchers to develop automatic decision support tools in order to provide specialists with a reliable instrument to identify PD. Various speech features have been proposed that are considered to provide relevant information about PD and help in the differentiation between speech of PD patients and that of healthy controls. In [8] a set of features, such as jitter, shimmer, harmonics to noise ratio and F0-related features, is assigned into different classes of dysphonia measures useful in replicating the clinicians’ scale (unified Parkinson’s disease rating scale, UPDRS) to assess the degree of PD severity.

The first main requirement for any speech analysis is the database of the investigated speech material. So far, the common speech type used in PD research has been sustained vowels. The choice is justified by the empirical evidence that speech signals associated to healthy vowel phonations do not show excessive variations along the production time (the signal is kept steady to certain extent). This type of speech material consists of recordings of vowels, pronounced for a given time duration (generally for some seconds). Such recordings are easy to use and robust; therefore they are common in clinical practice in any field of research [9]. However, examination of sustained vowels only relates to the operation of the vocal cords, and implies a steady state of the articulation organs. On the other hand, running speech gives a long time information on the complete movement of the articulation organ, containing words, sentences and paragraphs. Although there are surely features that characterize PD speech and are not present in production of sustained vowels (such as prosodic features), automatic processing (annotation, segmentation) of a long running speech (either spontaneous or elicited from read texts) is definitely harder and less robust. Speech recognition methods using forced alignment can be used to process read texts that contain (theoretically) the same language content, but manual correction of the resulted segmentation is necessary.
Spontaneous speech is even harder to process, and needs a lot of effort and human work. However, at least running speech consisting of read texts is sufficient to provide a full view of the acoustic features of PD.

In the literature, there are even some databases that contain running speech in different languages [10][11]. This will add variability to the extracted acoustic features that may differ from language to language. The research performed on this study is based on a Spanish database, thus its results can also be relevant in consideration of the language difference issue.

Decision support tools discriminate PD patients from healthy controls by means of automatic classification methods. Many classifiers are available with different distinctive strength. The most commonly used statistical classifiers [8][10][13][13] are Bayesian classifier, support vector machines, (deep) neural networks, Gaussian mixture models, random forests and k-nearest neighbours. In [8] about 97% classification performance was reported on the data consisting of sustained vowels for two classes: PD and healthy control. In another study [11] about 85% highest accuracy was achieved with speech intelligibility features from running speech using three classes: healthy, mild PD and severe PD.

The INTERSPEECH 2015 Special Session "Computational Paralinguistics Challenge (ComParE): Parkinson’s Condition Sub-Challenge" [16] calls researchers to apply acoustic feature extraction and machine learning methods to classify speech samples containing utterances from patients with Parkinson’s disease. The main task is to estimate the unified Parkinson’s disease rating scale (UPDRS scores) of each audio sample.

The present paper describes our method of dealing with the problem of classification. In Section 2 the most important characteristics of the released database are mentioned. Section 3 describes the method that is used to automatically segment and pre-process the speech samples. It also describes an applied database partitioning and automatic classifier (regression) method. Section 4 covers the results, and Section 5 is about the conclusions that can be drawn from the results.

2. Database

The database of the speech of patients with Parkinson’s disease was created by J. Orozco-Arrovave et al [1]. The samples that were made available for the challenge contain speech in Spanish from 50 people (25 male, 25 female) suffering from PD. The age of the male utterances ranges from 33 to 77 (mean 62.2 ± 11.2), the age of the female utterances ranges from 44 to 75 (mean 60.1 ± 7.8). The data comprises a total of 42 speech tasks per speaker, including 24 isolated words, 10 sentences, one reading text, one monologue, and the rapid repetition of syllables. The total duration of the recordings is 1.4 hours. The neurological state of the patients was evaluated by an expert neurologist according to the unified Parkinson’s disease rating scale (motor subscale): UPDRS-III. The values of the neurological evaluations performed over the patients range from 5 to 92. The audio files are divided into training and development sets. An additional test partition with labels unknown to the participants comprises a further eleven subjects. The complete description of the database can be found in [1].

3. Feature Extraction

This paper utilizes acoustic features different from the set of features that the organizers of the challenge made available.

3.1. Speech Segmentation

Rather than extracting features from the total duration of each sample, the computation was done at the middle of each vowel. Because no segmentation was available for the database, our language independent segmentation method was utilized in order to detect the middle of the vowels and the silence parts of the audio samples [14]. The speech was segmented into categories of speech sounds according to acoustic properties (middle-high vowel, low vowel, nasal, voiced and unvoiced bursts, etc.) using a support vector machine-based classification method [14].

3.2. Acoustic Features

The extracted acoustic features and their computational parameters are shown in Table 1. Each feature was calculated at the middle of the vowels in words, sentences and continuous speech, and was averaged for each audio sample. For each audio sample the mean and the variation of the features were calculated.

Table 1: Description of extracted features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature no.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter</td>
<td>1</td>
<td>F0 frequency variation over 3 F0 periods</td>
</tr>
<tr>
<td>Shimmer</td>
<td>1</td>
<td>Intensity variation over 3 F0 periods</td>
</tr>
<tr>
<td>Articulation rate</td>
<td>1</td>
<td>Frequency of syllables over the total duration of the audio sample</td>
</tr>
<tr>
<td>Intensity</td>
<td>1</td>
<td>Intensity of speech using 100 ms computational window</td>
</tr>
<tr>
<td>Intensity variation</td>
<td>1</td>
<td>Variation of intensity (with same computational parameters as for ‘intensity’)</td>
</tr>
<tr>
<td>Rate of transients</td>
<td>1</td>
<td>Ratio of transient parts of speech to the total duration of the audio sample [15]</td>
</tr>
<tr>
<td>Mel-frequency spectrum</td>
<td>27</td>
<td>Power Spectral Density using mel filterbank</td>
</tr>
</tbody>
</table>

3.3. Database Partitioning

Because the features that describe rhythm behave differently according to the duration of the produced speech and text material (long text, sentence, word, repetition of syllables) the database was partitioned into three parts: (1) long audio files lasting more than 10 seconds (texts, monologues); (2) medium audio files lasting between 2 and 10 seconds (sentences); and (3) short audio files with duration less than 2 seconds (word, repetition of syllables). The audio samples were used in the UPDRS score estimation first with the partitioning and then without, and the results then compared.
4. Classification method

The classification of the samples was done using linear regression by the method described in the following Section.

4.1. Simple Linear Regression

Because the number of samples representing each individual UPDRS score was low, the samples were merged according to their UPDRS scores by the following method. The total UPDRS score range was divided into ten subsets: 0-10, 11-20, 21-30, and so on up to 91-100. Each speech sample was assigned into the subset into which its original UPDRS score falls. During the training, the mean values of the subsets were used as UPDRS scores instead of the original ones. For example a sample with original UPDRS score 22 was used with UPDRS score 25 in the training. This method ensured a better statistical estimation of the mean values and variations for the acoustic features.

The mean acoustic features that were extracted for the audio samples were averaged over the UPDRS subsets. For each feature (for a given partition, see Section 3.3) there were ten pairs of UPDRS-extracted value (x, y) pairs, where x is the UPDRS score belonging to a given subset and y is the mean value of the extracted feature in the subset. A linear regression model was fit for each feature, both with partitioning and also without partitioning:

\[ y^k_{ij} = a^k_j + \beta^k_j x^k_i, \]

where y is the extracted value of sample i for the feature j in partition k, and \( x \) is the corresponding UPDRS score of sample \( i \) in partition \( k \). An example for the regression in the case of jitter and partition of long audio samples is shown on Figure 1.

![Figure 1: Linear regression model for jitter computed on the partition of audio samples with medium duration](image)

The procedure was performed for all partitions of the training samples (of long, medium and short duration) and also on the entire database without the partitioning (including all training samples).

4.2. UPDRS Score Calculation

For each audio sample multiple UPDRS scores were assigned according to the linear regression model of acoustic features. The final UPDRS score of a sample was obtained by a weighted sum of the partial scores:

\[ UPDRS^k_i = \sum_{j=1}^{n} UPDRS_{ij} \cdot w^k_j, \]

for sample \( i \) of partition \( k \), where \( j \) is the number of the feature and \( n \) is the total number of features. The weights of feature \( j \) and partition \( k \) were obtained by calculating the Spearman correlation for each individual feature on the training set. The correlations were normalized to sum up to 1 (\( \sum_{j=1}^{n} w^k_j = 1 \)) and were assigned as weights.

5. Results

The calculation of the final UPDRS scores was done for both the training (using 4-fold cross-validation) and the development set (using the entire training set for the training), both using database partitioning and also without using it. On the test set the detailed partitioning results are not available, only the result for the total test set could be uploaded, subsets were not accepted. Table 2 reports the results that were obtained for the different cases and also the baselines that are presented in [16]. In case of database partitioning the correlation value is calculated for the total database and also for the individual partitions.

However, it must be noted that it is uncertain which results should be included as comparison, because the parameters of the SVM in [16] giving the best result on the test set gave the worst result on the training and the development sets. Therefore, both the best values obtained for the ‘train CV’ and ‘train/devel’ sets (marked as baseline1) and the value that was obtained for the best parameters for the test set (marked as baseline2) are presented.

Table 2: Spearman correlation (\( \rho \)) obtained by linear regression. Baseline1 is the best baseline value for the given set (train CV and train/devel). Baseline2 is the correlation obtained with the best parameters for the test set. In case of database partitioning the correlation value is calculated for the total database and also for the individual partitions.

<table>
<thead>
<tr>
<th></th>
<th>train CV</th>
<th>train/devel</th>
<th>train/test</th>
</tr>
</thead>
<tbody>
<tr>
<td>with partitioning</td>
<td>.421</td>
<td>.425</td>
<td>.165</td>
</tr>
<tr>
<td>long</td>
<td>.589</td>
<td>.552</td>
<td>-</td>
</tr>
<tr>
<td>medium</td>
<td>.505</td>
<td>.334</td>
<td>-</td>
</tr>
<tr>
<td>short</td>
<td>.384</td>
<td>.464</td>
<td>-</td>
</tr>
<tr>
<td>without partitioning</td>
<td>.416</td>
<td>.347</td>
<td>.172</td>
</tr>
<tr>
<td>baseline1</td>
<td>.434</td>
<td>.492</td>
<td>.236</td>
</tr>
<tr>
<td>baseline2</td>
<td>.238</td>
<td>.368</td>
<td>.390</td>
</tr>
</tbody>
</table>

6. Discussion

The results show that although the best baseline correlation values in the case of train and development sets (baseline1) are still slightly higher than the results obtained with our method, the best values obtained by the SVM parameters most suitable for the test set (baseline2) are significantly lower than those achieved by the proposed regression method in the case of the training and development sets. However, in terms of the final results on the test set the performance rates are below baseline levels. This may be due to the high variation between the test and the train/devel speech material. Although the same test material was recorded, they were done in different sound environments. The baseline rates show the same phenomena: the lower the baseline values for the train/devel evaluation, the higher performance was achieved in the train/test case.

Except from the test set, the partitioning of the database according to sample durations clearly increased the results
(train/devel, $\rho=.425$ (with partitioning) and .347 (without partitioning)). This implies that the proposed features provide a better estimate of the UPDRS scores when different models are built for samples of different durations. Partitioning in the case of the test set has no effect on the already low correlation. This may also be due to the possibly high variation of the sound materials. Unfortunately, detailed partitioning result are not available due to the result submitting method of the challenge. The partitioning algorithm used in the paper is based on the duration information of the current database. A more complex and precise partitioning could be done applying ASR systems, discriminating the speech samples based on phoneme number, syllable number, word number, etc., but in the present study such tools were not available for the challenge.

The Spearman correlation values differ in the case of the three different partitions. The UPDRS scores of the long speech samples were the best estimated ones both for the training and development sets. This suggests that although the extracted features have large variations, averaging them over a sufficiently large time interval can provide a more precise estimate for UPDRS scores. The relatively high value of $\rho$ in the case of short samples (using development set as testing) may be due to the syllable repetition speech samples that may in turn be correlated with the severity of the Parkinson’s disease. This phenomenon may warrant deeper examination.

Table 3 shows the $\rho$ values for each feature individually using database partitioning on the ‘devel’ set. Articulation rate and rate of transient features (those that provide rhythmic information) behave differently in the partitions. Long audio samples correlate highly with these features due to the inclusion of possibly high rhythmic variation within their UPDRS scores. This implies that rhythm is an important measure of Parkinson’s disease (as former studies and the neurological background of the disease suggest, see Section 1). The role of articulation rate can also be seen in the case of the short audio samples, because they include fast repetition of syllables that also carry a high level of rhythmic information. On the other hand, rhythmic information of medium audio samples (mostly sentences) may have possibly high variation not only owing to UPDRS scores, but also because it is speaker dependent. Hence low discriminant strength is gained in this partition (low $\rho$ values).

Table 3: Spearman Correlation ($\rho$) obtained by linear regression for individual features for database partitioning (audio samples with long, medium and short duration).

<table>
<thead>
<tr>
<th>feature</th>
<th>partition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>long</td>
</tr>
<tr>
<td>jitter</td>
<td>.161</td>
</tr>
<tr>
<td>shimmer</td>
<td>-.012</td>
</tr>
<tr>
<td>articulation rate</td>
<td>.294</td>
</tr>
<tr>
<td>intensity</td>
<td>.114</td>
</tr>
<tr>
<td>intensity variation</td>
<td>.005</td>
</tr>
<tr>
<td>rate of transients</td>
<td>.322</td>
</tr>
</tbody>
</table>

Figure 2 shows the Spearman correlation values for the mel-band intensities. There are some highly apparent mel-bands that show relatively high correlation. This suggests that the corresponding frequency bands (around 200 Hz, 700 Hz, 1,500 Hz and 3,000 Hz) play a higher role in estimating the severity of Parkinson’s disease. Of course, this phenomenon may be language dependent, and these frequency bands might be characteristic for Spanish.

7. Conclusions

The paper describes our contribution in the INTERSPEECH 2015 Special Session “Computational Paralinguistics Challenge (ComParE): Parkinson’s Condition Sub-Challenge”. The main goal of the challenge is to perform automatic classification (regression) on speech of patients with Parkinson’s disease. By using linear regression models on the proposed acoustic features, and by partitioning the speech samples according to their total duration, we achieved lower Spearman correlation on the UPDRS scores provided we consider solely the best baseline result on the development set (‘baseline2’). The baseline results on the test sets outperformed our method, probably due to the high variation between the speech recordings of the test and the train/development sets. However, if we consider the final official baseline that belongs to the accuracy of the SVM with parameters of the best result on the test set, using the development set our obtained correlation is higher on the development set.

It must be emphasized that these correlations (and also the baseline results) count as weak. We experienced high intra-variation of the extracted features.

8. Acknowledgements

We would like to thank the organizers of the INTERSPEECH 2015 Special Session “Computational Paralinguistics Challenge (ComParE)” for making the database available and for preparing the challenge.

9. References


