Automatic Recognition of Unified Parkinson’s Disease Rating from Speech with Acoustic, i-Vector and Phonotactic Features

Guozhen An1, David Guy Brizan1, Min Ma1, Michelle Morales2, Ali Raza Syed1, Andrew Rosenberg3

1Department of Computer Science, CUNY Graduate Center, USA
2Department of Linguistics, CUNY Graduate Center, USA
3Department of Computer Science, Queens College (CUNY), USA
{gan, dbrizan, mma, mmorales, asyed2}@gradcenter.cuny.edu,
andrew@cs.qc.cuny.edu

Abstract

Parkinson’s Disease is a neurodegenerative disease affecting millions of people globally, most of whom present difficulties producing speech sounds. In this paper, we describe a system to identify the degree to which a person suffers from the disease. We use a number of automatic phone recognition-based features and augment these with i-vector features and utterance-level acoustic aggregations. On the Interspeech 2015 ComParE challenge corpus, we find that these features allow for prediction well above the challenge baseline, particularly under cross-validation evaluation.

Index Terms: Parkinson’s Disease Recognition, Phonotactic Modeling, Computational Paralinguistics

1. Introduction

Named after the 19th century surgeon James Parkinson, Parkinson’s disease (PD) is a progressive disease which produces a number of sensorimotor symptoms including tremors while at rest, rigidity, akinesia (impairment of voluntary motor control), and postural instability [1], i.e. lack of balance. Some of these “TRAP” symptoms may lead to speech problems or speech dysfunction. In addition to these symptoms, the disease has also been shown to be responsible for a number of neurological and psychological symptoms such as changes in sleep patterns, loss of emotional well being, cognition, visual and spatial deficits, as well as changes in perception [2].

After Alzheimer’s Disease, PD is the second most common neurodegenerative disorder, affecting about 1% of people over the age of 65 [3]. Rates of incidence of PD range from 12 to 15 cases per 100,000 people in Europe and the United States— including actor Michael J. Fox and former heavyweight boxing champion Muhammad Ali— but these rates increase with age and change with gender, ethnicity and possibly with socioeconomic factors [4].

There is no known cure for PD; however, there are treatments and therapies to mitigate the effects of the disease [3]. In particular, speech therapy has been shown to be effective under the guidance of trained counselors [5]. We are interested in building automatic speech-based tools to assess the effects of PD, specifically to determine the effectiveness of speech therapy along with other treatments and therapies for this disease.

In this paper, we describe a two-stage approach to detecting Parkinson’s Disease from the speech signal of an affected subject. This work describes a submission to the Interspeech 2015 ComParE Parkinson’s Condition Sub-Challenge [6]. In this challenge, a speaker’s utterances are given a label on the Unified Parkinson’s Disease Rating Scale (UPDRS). Following the baseline classification, we add features useful in predicting the subjects’ scores on the UPDRS, which include syllable duration, emotion and phonotactics. We find that incorporating these features results in a Spearman correlation with human UPDRS values of 0.2761, an absolute increase of 0.0401 over the challenge baseline.

2. Related Work

Perhaps because PD is a disease affecting so many people, a substantial amount of research has been conducted on PD symptoms. Dysarthria, speech disturbances caused by the neurological portion of the disease, has been the focus of most research due to the large percentage of PD patients it affects, as high as 90% [7]. For example, Scott and Caird [5] find that the main features of speech disorder of PD patients are reduced intensity of voice, a tendency toward overall increased pitch, monotony of speech, and an abnormal rate of speaking. Skodda et al. [8] build on this by focusing on the stability and duration measures applied to syllables /pal/ and /bal/. Operating under the loss of control hypothesis, Walsh and Smith (2011) [7] find higher rates of disfluency for longer or more syntactically complicated utterances in mild-moderate PD subjects, with implications for patients with more severe symptoms.

Ho et al. [9] look at similar voice qualities predictive of PD, such as harshness, reduced volume, disturbed intonation, fluency, inappropriate pausing, syllable repetition and imprecise articulation resulting from lack of motor control. The study classifies speech impairment in 200 patients with PD into five levels of overall impairment severity and describes the corresponding type (voice, articulation, fluency). Most (73.5%) of the subjects in the study demonstrate a gradual deterioration of speech features, almost always involving voice first, before progressing to the prominent voice and articulation pattern, with the latter being the most severely affected.

In more recent work, Skodda et al. (2011) [10] and Skodda et al. (2012) [11] analyze the loss of articulation capacity by using different measures based on 2 formants (F1 and F2) of the voice spectrum: VAI (vowel articulation index) and tVSA (triangular vowel space area).

The Unified Parkinson’s Disease Rating Scale (UPDRS) was introduced in 1987 by Fahn and Elton; it enjoys popularity as an instrument to quantify the longitudinal effects of the disease or therapies on patients [12]. In associating a speech signal to the UPDRS, Orozco-Arroye et al. [3] focus on 5 vow-
els of Spanish-speaking subjects, looking at a number of noise measures, as well as periodicity and stability features such as jitter. Their results support that variability of pitch is a good cue in characterizing vowels uttered by people with a PD diagnosis.

While much of the research is focused on the articulatory difficulties exhibited by PD patients, Lieberman et al. [13] find significant correlation between an increase in syntactic errors on a language assessment task and voice onset time (VOT) errors on a single-word production task. Errors in comprehension of syntax are measured by RITLS (Rhode Island Test of Sentence Comprehension).

For our work on the Sub-Challenge, the data are drawn from the speech of 50 individuals with PD and 50 control subjects with no known PD diagnosis. All subjects are native Colombian adults and all native speakers of Spanish. Further details of the corpus and stimuli appear in [3].

3. Method

In this section, we outline our approach in two types of experiments: 1) Addition of Parkinson-detecting features (cf. Section 3.1); and 2) Incorporation of Total Variability (cf. Section 3.2).

Each set of experiments uses the same framework; as a part of the Interspeech 2015 ComParE challenge, all experiments are performed using the challenge train, development and test sets of the Corpus. In the experiments we describe in this section, we estimate performance of our models using four-fold cross-validation on the train data. The features from successful experiments are used to train our models on the training data and evaluate the performance on the development data. Finally, we test the performance of our models using the test data.

The initial hypotheses used for all experiments are the baseline challenge results. These are generated using the baseline feature set and SMO (SVM) classification with C=0.001.

The challenge baseline was created with features extracted from a pre-release version of OpenSMILE [14], specifically version 2.1. In an attempt to recreate the baseline results, we performed our own feature extraction with OpenSMILE version 2.0, the most recent version available to us. In addition to the differences in software, we were unable to obtain the configuration for the initial hypothesis, so we used a set of features from the ComParE 2012 Speaker Trait Challenge (Pathology Sub-Challenge) [15], which we believe to be closely related. We report the performance resulting from these features and executed using this version of software as “CUNY Baseline” (BF) in our results. The feature set is specifically drawn from the Pathology Sub-Challenge. Using the the NKI CCRT Speech Corpus (NCSC), the task for the sub-challenge was to determine the intelligibility of Dutch speech from subjects who have had concomitant chemo-radiation treatment (CCRT) and as a result experience speech difficulty. While we do not equate PD with CCRT pathology, we believe that the underlying features may be useful in forming predictions on the UPDRS scale. See, for example, work performed in the 2012 and 2013 ComParE challenges which used a fusion of low-level and phonotactic features combined with machine learning techniques for predicting the intelligibility of speech [16] and the emotion with which the speech was delivered [17].

The CUNY Baseline feature set (or ComParE 2012 Speaker Trait Challenge feature set) contains a set of 4 energy-related low-level descriptions (LLDs), 54 spectral-related LLDs, and 6 voicing-related LLDs computed from the speech signal at the frame level. We include statistical functionals such as mean, maximum, minimum, standard deviation as well as percentiles applied to these frame-based LLDs to compute overall features. We add these to the BF’ baseline. While a number of the features overlap with the initial hypothesis, and this feature set yields 6,373 features—the same number as is provided by the official challenge baseline feature set—the sets are non-identical. Where the features are the same, many values are different due to variations in settings.

3.1. PD Features

Following the work of others, we add four sets of features to the baseline which we believed to have positive effect on the baseline features. Each feature set is described in this section.

3.1.1. Syllable-level Features

Studies by McRae et al. [18] and others show that speaking rate is negatively affected by level of PD severity. We measure speaking rate by calculating the ratio of syllables to the utterance duration—i.e. syllables per second. We augment our baseline with a number of syllable features.

We detect the pseudosyllable regions based on the Villing envelope based approach [19] as implemented in AuToBI [20], and derive the following features for each utterance: 1) total number of syllables, N; 2) total duration of syllable regions, \( \text{sum(syl)} \); 3) syllable ratio defined as the ratio of \( \text{sum(syl)} \) to \( \text{sum(utt)} \); 4) speaking rate on the syllable level, namely, \( \frac{\text{sum(syl)}}{N} \); 5) average duration per syllable, namely, \( \frac{\text{sum(utt)}}{\text{sum(syl)}} \); 6) the maximum, minimum, range, standard deviation, mean and median values of the duration of the syllables within each utterance. All the duration values used in 1) to 5) are raw time duration, while the duration values used in 6) is normalized by \( \text{sum(utt)} \).

3.1.2. Low-Level Descriptor (LLD) Features

Since there have been differences in baseline features sets over the years of the Interspeech Paralinguistics Challenge, we evaluate the feature sets from prior years on this task. The 2010 Affect Sub-Challenge (“emotion2010”) [21] was designed around recognizing emotion in speech. This feature set contains 34 low-level descriptors (LLD) with 34 corresponding delta coefficients and 21 functionals applied to each of these 68 LLD contours. To these, 19 functionals are applied to the 4 pitch-based LLD and their four delta contours are added. Finally, the feature set includes the pitch onsets of automatically recognized pseudo-syllables, and the total duration of the input. We normalize the absolute positions of maximum and minimum values in frames according to the segment length. Since some of the features in the 2010 Challenge feature set have been included in the baseline feature set, we prune this LLD set to 1,399 additional unique features.

3.1.3. Formant Features

The most relevant acoustic parameters for production of vowels are the frequencies of the first two formants, F1 and F2. Sapir et al. [22] hypothesize that these formant frequencies change as a function of the movements of speech articulators these parameters are distinctive in speech production affected by PD. For example, to form the vowel ‘i’, the frequency of F2 increases and that of F1 decreases as the tongue moves forward [23]. We use PRAAT to extract formant features of F1 and F2 at the utterance level [24].
3.1.4. Phonotactic Features

Due to loss of motor control, we hypothesize that PD patients may be distinguished by extended syllable durations and repetitions, assuming the syllable is unstable during articulation. We therefore add a number of phonotactic features (the distribution of monophones, biphones and triphones) and phone duration features to BF'. While this technique has been shown to be effective in distinguishing languages [25] and dialects [26] from each other, our premise is that it may also be useful in predicting the level of those with PD diagnosis.

For this experiment, we use phnrec [27] to extract the phone hypotheses and their durations in 4 languages (Czech, English, Hungarian, Russian), yielding over 31,000 new features. We add these phonotactic features to BF'.

3.2. i-Vector Approaches

Motivated by Joint Factor Analysis [28, 29], i-vector modeling was originally proposed in [30]. Later work in classifying cognitive load, showed the success of implementing an i-vector framework. The i-vector exploits the concept of “total variability”. It improves upon factor analysis by estimating a single low-dimensional subspace (i-vector space) where all variability is modeled, leading to improved accuracy and reduced computational complexity. Previous work exploited this framework to model speaker-specific variability across different levels of cognitive load; we develop two techniques based on i-vectors to model the variability across different severity levels of PD. In both cases, we use the open source tool ALIZE [31] as the basis of our framework.

3.2.1. Feature-selected i-vectors

We extract Mel-frequency cepstral coefficient (MFCCs) features and use this feature set to train our Universal Background Model (UBM), using all available training and development data. We then perform Best First feature selection to determine the highest-performing subset of the 10-best features. This subset is the basis of our i-vector features, reported in the section below. We determine the performance of these i-vector features followed by feature selection with and without the other features described in this section.

3.2.2. Stacked LDA Transforms

For our second approach with i-vectors, we explore an i-vector pipeline, similar to that in speaker recognition, with extraction and post-processing transformations, before performing any regression. While we do not equate speaker recognition with PD, we believe this approach may capture speech abstractions useful for PD diagnoses. Here, we experiment with higher dimensional i-vectors. For extraction, our final solution is based on a 512-mixture UBM used to learn a 200-rank Total Variability matrix, yielding 200-dimensional i-vectors.

A common step in an i-vector pipeline is to perform post-processing to compensate for channel variations and improve speaker representations. Typically, transforming i-vectors based on Linear Discriminant Analysis (LDA) has yielded good results in the speaker recognition tasks [32]. We conduct a number of experiments and settle on a two-layer stack of transformations produced by LDA variants: Probabilistic Linear Discriminant Analysis (PLDA) [33] followed by Local Fisher Discriminant Analysis (LFDA) [34].

PLDA is a supervised, generative model with two components for speaker and channel variation. The speaker-specific component models inter-speaker variation and is independent of each segment, while the channel-specific component of the model is dependent on each segment. Since the training data contains 27 distinct UPDRS values, we tried to capture Parkinson related speaker characteristics by using 27 corresponding classes for the speaker specific component model.

Parkinson’s related speech dysfunction may arise from different symptoms, even for patients within the same UPDRS. Further, speaker variation still exists within each class due to multiple patients. Thus we expect multiple clusters, rather than a coherent cluster, per class in our i-vector space. To mitigate this issue, we incorporate a second transformation layer, which does not require multimodal samples to fall into a single cluster. LFDA is a supervised metric learning technique which combines the benefits of LDA with Locality Preserving Projection [35]. It learns a projection to maximize between-class separability while preserving within-class local structure by keeping in-class data pairs close and between-class data pairs apart.

The UPDRS training labels did not cover all expected values and the data was not well balanced within UPDRS values. We perform a more coarse division of the training data into five quantities based on UPDRS intervals: [0, 19], [19, 30], [30, 40], [40, 52], [52, 108]. This choice was driven by the 0-4 ratings in the MDS-UPDRS with interpretations ranging from normal to severe along with an expectation that human diagnosis may be driven by similar coarse divisions [36]. The projected 200-dimensional i-vectors (training data) from the first layer are used to train the second LFDA layer based on these five classes and learn a second projection matrix. The two projections are then applied to development data.

4. Results

We experiment with several different configurations of features listed in the Methods section (cf. Section 3). First, we use four-fold cross validation on the training data for verification of the technique. Promising configurations are advanced to the second phase, in which we use models trained on the train data and evaluated their performance on the development data. We train a final subset of these models on the training and development data and evaluate their performance on the test data. We report the results of each phase in this section. For each set of results, we report the Spearman correlation coefficient.

4.1. Four-fold Cross-Validation

We use the Weka [37] tool to generate folds and results on the training data. We exclude all features with string attributes, use SMO Regression and keep all other values at their default settings. While this generates folds different from the challenge baseline [6], this enabled more rapid evaluation of the feature sets.

The first section of Table 1 summarizes the baseline (BF) Spearman correlation coefficient results for cross-validation on the training data and our attempt to recreate the baseline (BF') with the feature set and software as described in the previous section. The second section shows the effect of augmenting our baseline with the feature sets described in the previous section. We note that the performance of the feature sets largely improves system performance. The last section of this table shows that combining many of these feature sets leads to further performance boosts.

We observe that phonotactic features have the largest single positive impact over the BF' baseline. On the other hand,
Table 1: Cross-Validated Results on Training Data.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Spearman</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Features (BF)</td>
<td>0.6176</td>
<td></td>
</tr>
<tr>
<td>CUNY Baseline (BF')</td>
<td>0.6093</td>
<td></td>
</tr>
<tr>
<td>Syllable + BF'</td>
<td>0.6104</td>
<td>+0.0011</td>
</tr>
<tr>
<td>LLD + BF'</td>
<td>0.6586</td>
<td>+0.0493</td>
</tr>
<tr>
<td>F1 Formant + BF'</td>
<td>0.5042</td>
<td>-0.1051</td>
</tr>
<tr>
<td>F2 Formant + BF'</td>
<td>0.5040</td>
<td>-0.1053</td>
</tr>
<tr>
<td>Phonotactic + BF'</td>
<td>0.6739</td>
<td>+0.0646</td>
</tr>
<tr>
<td>i-vector + BF'</td>
<td>0.6104</td>
<td>+0.0011</td>
</tr>
<tr>
<td>Stacked LDA + BF'</td>
<td>0.6369</td>
<td>+0.0300</td>
</tr>
<tr>
<td>Syll., LLD + BF'</td>
<td>0.6594</td>
<td>+0.0418</td>
</tr>
<tr>
<td>i-vector, Syll., LLD + BF'</td>
<td>0.6607</td>
<td>+0.0501</td>
</tr>
<tr>
<td>S.LDA, i-vector, Syll., LLD + BF'</td>
<td>0.6937</td>
<td>+0.0844</td>
</tr>
<tr>
<td>All features + BF'</td>
<td>0.7088</td>
<td>+0.0995</td>
</tr>
</tbody>
</table>

the formant features individually have a negative impact on the baseline, so we exclude them from future experiments. Combining all features yields the highest performance, but since the combination is clearly non-additive, many of the features may capture similar phenomena with respect to UPDRS.

4.2. Development Results

In the second phase of experiments, we further evaluate the promising feature sets from the previous phase. These feature sets are trained on all the training data and evaluated on the development data. The results appear in Table 2.

Table 2: Experimental Results on Development Data.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Spearman</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Features (BF)</td>
<td>0.4915</td>
<td></td>
</tr>
<tr>
<td>CUNY Baseline (BF')</td>
<td>0.5076</td>
<td></td>
</tr>
<tr>
<td>Syllable + BF'</td>
<td>0.5083</td>
<td>+0.0007</td>
</tr>
<tr>
<td>i-vector + BF'</td>
<td>0.5086</td>
<td>+0.0010</td>
</tr>
<tr>
<td>LLD + BF'</td>
<td>0.5139</td>
<td>+0.0063</td>
</tr>
<tr>
<td>Phonotactic + BF'</td>
<td>0.4808</td>
<td>-0.0268</td>
</tr>
<tr>
<td>Stacked LDA + BF'</td>
<td>0.5207</td>
<td>+0.0131</td>
</tr>
<tr>
<td>Syll., LLD + BF'</td>
<td>0.5140</td>
<td>+0.0064</td>
</tr>
<tr>
<td>i-vector, Syll., LLD + BF'</td>
<td>0.5150</td>
<td>+0.0074</td>
</tr>
<tr>
<td>S.LDA + i-vector, Syll., LLD + BF'</td>
<td>0.5258</td>
<td>+0.0182</td>
</tr>
<tr>
<td>All features + BF'</td>
<td>0.5051</td>
<td>-0.0025</td>
</tr>
</tbody>
</table>

We note that some of the performance gains we see from the previous phase are not realized in this experiment. We attribute this to the difference in methodology; specifically, the development data was selected so that there is no overlap of speakers from the training data. As such, we believe that the phonotactic features, while very useful in predicting a UPDRS rating, model each speaker too closely. We therefore exclude them from the next phase of experiments.

4.3. Test Results

Before our final phase of experimentation, we first tuned the C parameter of SMO regression by experimenting with the baseline feature sets from the previous phase. This is the procedure we follow, reporting results in Figure 1 for this tuning on the development data only. We note the similarity of the curves between the official challenge baseline (BF) and our attempts to recreate the baseline (BF'). We use the output of this tuning, with C=0.001 for our test result baselines and experiments.

For the final phase of experiments, we evaluate the most promising models from the previous phase on the test data. With the exception of the C parameter, we leave all other parameters at their default values. The Spearman correlation coefficient results appear in Table 3.

Table 3: Experimental Results on Test Data.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Spearman</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Features (BF)</td>
<td>0.2360</td>
<td></td>
</tr>
<tr>
<td>CUNY Baseline (BF')</td>
<td>0.2605</td>
<td></td>
</tr>
<tr>
<td>LLD + BF'</td>
<td>0.2739</td>
<td>+0.0134</td>
</tr>
<tr>
<td>i-vector, Syll., LLD + BF'</td>
<td>0.2739</td>
<td>+0.0134</td>
</tr>
<tr>
<td>S.LDA, i-vector, Syll., LLD + BF'</td>
<td>0.2761</td>
<td>+0.0156</td>
</tr>
<tr>
<td>All features + BF'</td>
<td>0.2608</td>
<td>+0.0003</td>
</tr>
</tbody>
</table>

We see a carry-over of the gains observed when evaluating on the development data: each feature set has a positive impact on the baseline, with our best combination of systems performing with a Spearman correlation coefficient 0.0156 above the CUNY baseline (BF') and 0.0401 over the challenge baseline.

5. Conclusion and Future Work

In this paper, we explored a number of features motivated by research on Parkinson’s Disease to improve the automatic prediction of UPDRS scores from speech. This work was undertaken as part of the Interspeech 2015 Paralinguistics Challenge. We find an improvement of 0.0401 over the challenge baseline. This improvement is attributed to incorporating features derived from automatic syllable detection, effective post-processing of i-vector features. We had some promising results in incorporating phonotactic features, which ultimately modeled speaker differences too closely. Future work will develop more robust representations of phonotactics and speaking rate to improve speaker-independent modeling.

6. Acknowledgments

Hernisa Kacorri developed the Stacked LDA Transforms approach with author Ali Raza Syed. Without her efforts, this section would have been absent from our work. The authors thank her immensely for her contributions.
7. References


