Random Forest-Based Prediction of Parkinson's Disease Progression Using Acoustic, ASR and Intelligibility Features

Alexander Zlotnik[1], *Juan M. Montero*[1], *Rubén San-Segundo*[1], *Ascensión Gallardo-Antolín*[2]

[1]Speech Technology Group, ETSIT, Universidad Politécnica de Madrid, Spain
[2]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain

azlotnik@die.upm.es, juancho@die.upm.es, ruben.sansegundo@upm.es, gallardo@tsc.uc3m.es

Abstract

The Interspeech ComParE 2015 PC Sub-Challenge consists of automatically determining the degree of Parkinson's condition using exclusively the patient's voice. In this paper, we face this problem as a regression task and in order to succeed, we propose the use of an ensemble learning method, Random Forest (RF), in combination with features of different nature: acoustic characteristics, features derived from the output of an Automatic Speech Recognition system (ASR) and non-intrusive intelligibility measures. The system outperforms the baseline results achieving a relative improvement higher than 19% in the development set.

Index Terms: random forest, regression, Parkinson's disease, ASR features, intelligibility

1. Introduction

Parkinson's Disease (PD) is a chronic progressive neurodegenerative disorder that affects an estimated seven to ten million people worldwide, increasing its incidence with age. Persons with PD may experiment with different levels of severity the following motor symptoms: resting tremor, bradykinesia or slow movement, rigidity and postural instability. Other non-motor symptoms, as disorders of mood, behaviour and cognition and alteration of speech, are also common.

As PD is a progressive disease, it is very important to track its symptom progression, which is currently monitored by the Unified Parkinson's Disease Rating Scale (UPDRS). UPDRS comprises three components, namely, mentation, behaviour and mood; activities of daily living; and motor. The monitoring of PD progression is performed by expert medical staff, is costly and requires the physical presence of the patient in the clinic, which is sometimes troublesome. In the last years, several attempts have been carried out for finding solutions to these problems. Among them, the use of speech tests have attracted the attention of numerous researchers, as it is an non-invasive and fast method which could allow to perform remote monitoring of patients and feedback in their voice treatment.

For these reasons, the task of determining the degree of Parkinsons condition using only the information contained in patients' voice is nowadays a relevant challenge. In this context, the Parkinson's Condition (PC) Sub-Challenge, of the Computational Paralinguistics Challenge (ComParE), Interspeech 2015 [1], addresses this issue.

According to recent studies, approximately 70%-90% of patients with PD show some form of vocal impairment [2]. Besides, it seems that there is a strong relationship between PD progression and speech degradation [3]. It is worth mentioning that speech is explicitly assessed in the last two components of UPDRS in order to subjectively measure its intelligibility and its expressivity.

Speech problems due to PD are produced by incoordination or reduced movements of the muscles involved in breathing and voice production mechanisms. As a consequence, PD patients may suffer phonatary and articulatory impairments (hypophony or tremulousness), as well as unnatural prosody due to alterations in rhythm, intonation and speaking rate. In fact, persons with PD usually tend to speak softly and slowly (although in other cases, speech may become faster) and experiment difficulties beginning sentences yielding, in some occasions, hesitation and/or repetitions of words or sounds. All these effects turn into a decrease in intelligibility.

First studies dealing with the problem of automatically tracking the PD progression through speech, used only sustained vowel recordings from which a set of dysphonia measurements were extracted [4], [5]. These features (jitter, shimmer, harmonics-to-noise ratio, etc.) are mainly related to phonation and articulation. However, as pointed in [6], the analysis of prosody and intelligibility requires the inclusion of running speech. In this context, in the work by [7], in which pitch-related cues were also considered, it is shown that a reading task is better for automatically assessing the severity of PD than sustained phonation tasks.

Our approach to the PC subchallenge goes into two main directions. In the first one, we propose the use of features which represent speech aspects different to those covered by the baseline acoustic characteristics, as for example, rhythm and tempo and some measures of the voice intelligibility. Related to this last point, we describe also several features extracted from the output of an automatic speech recognition applied over the PD speech. The second line refers to the study of different learning algorithms for regression, like Ranking SVM and Random Forests, which has been used for ranking tasks (like information retrieval). We hypothesize that they may provide good performance for UPDRS prediction as Spearman correlation is the official evaluation measure of the subchallenge.

The remainder of this paper is organized as follows: Sections 2 and 3 describe, respectively, the different feature sets and regression techniques considered for the PC task. Our results are presented in Section 4, followed by some conclusions of the research in Section 5.

2. Feature Extraction

In this section, we describe the different features we have considered for the PC task. These characteristics are related to the four dimensions in which speech is affected and distorted by the Parkinson's Disease: phonation, articulation, prosody and

intelligibility.

2.1. Preprocessing and Feature Normalization

Silence regions of the audio signals do not contain phonetic and/or articulatory information of speech and, however, they can distort the statistics and other measures computed over audio segments (or the whole utterance) of low-level descriptors. For this reason, as a preprocessing stage, frames with energy below 5% of the median of the energy of the whole utterance were removed. Elimination of low energy frames was only performed prior to the extraction of the baseline acoustic features (see below) and not for MIR-based, intelligibility and ASR-based characteristics.

Once training, development and test features are extracted, they are normalized by subtracting the mean and dividing by the standard deviation computed over the training data.

2.2. Baseline Acoustic Features

Two baseline acoustic features are considered. The first set corresponds to the one proposed in the PC Subchallenge [1]. They have been extracted using the feature extraction toolbox openSMILE version 2.1 [8] and consist of a set of 6373 utterance-level acoustic features computed over low-level descriptors such as energy, Mel-Frequency Cepstrum Coefficients (MFCC), pitch and voice quality features as jitter, shimmer and harmonic-to-noise ratio (HNR), etc. The second set is a slight modification of the first one with the difference that 18 MFCC (instead of 14) are computed. Note that these features cover phonetic, articulatory and prosodic characteristics of the impaired speech.

2.3. MIR Features

As speech of persons with PD suffers from rhythmic inadequacies and other sound distortions problems, we decided to extract some features related to these characteristics by considering the speech prosody as a kind of musical sound with a typical cadence. Before some preliminary experimentation with different features coming from the field of Music Information Retrieval (MIR), we finally chose the following ones:

- *Beatspectrum*. It is a measure of acoustic self-similarity as a function of time lag, which is useful for the characterization of the rhythm and tempo of musical recordings. It has been computed using the method proposed in [9].
- *Roughness*. Also called sensory dissonance, it measures the beating phenomenon when pair of sinusoids are closed in frequency. It is estimated by computing the peaks of the spectrum and taking the average of all the dissonance between all possible pairs of peaks [10].
- *Spectral Irregularity*. It represents the variability between adjacent peaks of the spectrum and it has been calculated using the formulation described in [11].

All these features have been extracted using the MIRtoolbox ver. 1.6.1 [12],

2.4. ASR Features

The use of an Automatic Speech Recognizer could improve the performance of a Parkinson progression predictor, assuming that speech progressively affected by the disease should decrease the performance of the ASR system. Given the database comprises Latin American speakers, we have tried two approaches:

- Standard context–independent HTK-based Spanish phoneme and syllable recognizers, trained on a Castilian Spanish database (Albayzin, available at ELRA [13]).
- the Google Speech Recognition API (by means of the free SpeechRecognition 1.1.4 python package), available not only for Castilian Spanish but also for Colombian Spanish. The output of the recognizer is a list of word recognised sequences and a confidence measure.

From the phoneme string of the first recogniser, we can compute new features to be used for Parkinson prediction:

- number of phonemes recognised
- average likelihood per phoneme
- average speech rate (phonemes)
- maximum phoneme likelihood
- minimum phoneme likelihood
- number of syllables recognised
- average likelihood per syllable
- average speech rate (syllables)
- maximum syllable likelihood
- minimum syllable likelihood

From Google ASR output, we can compute new features too:

- was the recognition successful?
- number of phonemes recognised
- highest confidence measure
- average number of recognised vowels
- average speech rate

2.5. Intelligibility Features

As mentioned before, in most of the cases, the speech intelligibility of persons with PD degrades according to the level of severity of the disease. For this reason, we decide to study the possibility of including some measures related to intelligibility in the feature set. Objective methods for automatically predicting speech intelligibility can be classified into two groups: intrusive and non-intrusive. Whereas in the first class, a "clean" reference signal is needed in order to calculate its distance to the signal to be assessed, in the second case, no reference signal is required. Most of the conventional objective measurement methods belong to the first group. However, in our case, as clean reference signals are not available, it forced us to adopt non-intrusive techniques. In particular, we have considered two different methods, as described in next subsections.

2.5.1. SRMR Features

The measure named Speech to Reverberation Modulation energy Ratio (SRMR) was originally proposed for intelligibility estimation of (de)reverberant speech [14]. SRMR is based in the observation that, for clean speech, the modulation energy is mainly located at modulation frequencies below 20 Hz with maximum values around 4 Hz (rate of syllables), while higher modulation frequencies are mainly due to reverberation. Following this idea, SRMR is basically defined as the ratio between

the average modulation energy below 20 Hz and above this frequency.

We hypothesize that, although the voice of PD patients is not affected by reverberation (if recorded in clean conditions), the artifacts due to the voice impairments could produce some energy regions in high modulation frequencies. Following this idea, we decided to use the values of SRMR computed over the whole utterance as an additional feature.

2.5.2. P563 Features

The second set of features related to intelligibility were based on the ITU-T standard P.563 [15]. This standard computes a non-intrusive measure of quality as a linear combination of several signal parameters which have been designed for detecting the following main distortions: unnaturalness of speech, strong additional noise and interruptions, mutes and time clipping. Among the 51 signal characteristics extracted by the standard, we have used only those related to the vocal tract analysis carried out for measuring the degree of unnaturalness, which were not already included in the baseline acoustic set. In particular, these features are:

- *VTPMaxTubeSection*: For computing this parameter, the human vocal tract is modelled as a set of tubes of different lengths and time varying cross-sectional areas. In particular, *VTPMaxTubeSection* is the maximum section size of the first tube over the whole input signal.
- *FinalVtpAverage*: As in the previous case, after the human vocal tract modelling, *FinalVtpAverage* represents the averaged section of the last tube.
- *BasicVoiceQuality*: This value represents an estimate of the audible disturbance. More details about its computation can be found in [15].

3. Regression Techniques

3.1. SVR and Ranking SVM

Support Vector Machines (SVM) have been proved to be very good for supervised classification [16]. In addition, SVM can be extended to solve regression tasks [17]. In this case, the method is called Support Vector Regression (SVR).

As the evaluation measure of the UPDRS prediction task is the Spearman correlation, actually, we are not very interesting in estimating the exact values of UPDRS but their relative ranking position with respect to the reference UPDRS labels. For this reason, it was considered the use of a modification of the traditional SVM/SVR for ranking purposes, called Ranking SVM. The idea behind Ranking SVM is that, instead of using SVM for regression, the ranking optimization problem becomes equivalent to that of classifying SVM on pairwise difference vectors.

A good implementation of this strategy is SVM-rank toolkit [18], [19].

3.2. Random Forests

A Random Forest (RF) is an ensemble of regression trees trained for randomly-sampled sets of data, sharing a common distribution [20]. The algorithm first creates a "bag" of samples by random sampling from the training set; then creates a tree-based ranker for each "bag" of data; and finally ensembles the full forest of trees.

In this paper we have used an open-source implementation of Random Forest, RankLib 2.1, included in the Lemur project [21], [22]. RankLib package also contains other ranking-oriented algorithms such as: RankNet, RankBoost, AdaRank, Coordinate Ascent, LambdaMART, ListNet.

However, in the experiments of this paper, the best results have been obtained by means of Random Forests using Multiple Additive Regression Trees (MART).

In ranking-based problems such as this challenge's, instead of using information gain measures and maximum-likelihood estimates, listwise metrics from Information Retrieval techniques should be used:

- Normalized Discounted Cumulative Gain cut at the top 10 elements (NDCG@10) [23]
- Expected Reciprocal Rank at the top 10 elements (ERR@10) [24]

However, the RankLib implementation we used only allows using MAR Trees, also known as Gradient boosted regression trees [25] as the basic bag ranker. MART only allows Root Mean Square Error pair-wise loss as the optimization criteria.

The parameters of the training process that can be controlled are:

- the number of bags
- the sub-sampling rate
- the feature sampling rate
- the number of trees in each bag
- the number of leaves for each tree
- the shrinkage or learning rate
- the number of threshold feature candidates for tree splitting.
- the minimum number of samples each leaf has to contain

The file format of the training, validation and test files is the same as for SVM-Rank (compatible with the libsvm format).

4. Experiments

The experiments have been performed using the database described in [6] and provided for the PC Sub-Challenge [1]. It consists of 50 speakers for training and development and 11 additional subjects for test. The language is Spanish. It is worth mentioning that there is a mismatch between the acoustic environment in which the training/development files and the test files were recorded making the task even more difficult.

Results are given in terms of the Spearman Correlation (measure of the competition) between the UPDRS values for each patient determined by a neurologist expert and the predicted UPDRS values by the different learning algorithms. Table 1 shows the results achieved by different feature sets and the SVR learning algorithm.

Set of features of the sequence of experiments in Table 1:

- Feat1: Acoustic baseline features (ComparE)
- Feat2: ComparE with 18 MFCC and suppresion of low-energy frames
- Feat3: Feat2 + Feature Selection performed by Elastic Net
- Feat4: SRMR + MIR + P563 features
- Feat5: Feat3 + Feat4

Table 1: *Spearman Correlations for different feature sets and SVR*

Features	SVR
Feat1	0.4920
Feat2	0.5353
Feat3	0.5572
Feat4	0.3628
Feat5	0.5629

4.1. Baseline results

The speaker–independent baseline result (Feat1) on the development set is 0.492 (Spearman correlation) with standardized input features. This result was obtained by using all the features provided by the challenge and the Weka's SVR implementation with SMO optimization with the default configuration (C=0.001 and L= 1.0).

By optimizing C and L, only a slight improvement is obtained on the development set (0.498, C=0.001, L=0.8).

4.2. Basic feature selection

Some standard feature preprocessing and selection procedures have not been successful. Neither Weka's feature selection nor PCA nor MRMR (Minimum-Redundancy Maximum-Relevancy) have been able to improve the baseline results when using Weka's SVR.

4.3. ASR and acoustic pre-processing

The combination of the full set of original features and the features from the ASR (phoneme and syllable recognisers) into a bigger set has not been successful.

However, by removing the silence segments in the original audio files and re-computing the features file with the openSMILE package (including 18 MFCC instead of just 14), the baseline is significantly beaten: 0.513. By optimizing the C and L parameters, one can get up to 0.535 (Feat2).

4.4. Elastic Net feature selection and ASR

By applying Elastic Net, a Lasso extension, with gamma = 0.025 and l1_ratio = 0.7, after removing silences, the set of 6773 features can be reduced to just 427 non-zero features. With this reduced set of features SVR (C = 0.3, epsilon = 1 and gamma = 0.01), can achieve 0.557 on the development set (Feat3).

Although ASR features were not good enough when combined with the original set of features, we can try to improve the output of this last 0.557 experiment by means of a post-regressor (we have tried several kind of regressors, but the best result was obtained with a linear Weka's implementation). The combination of this output and features from ASR, gets 0.569.

4.5. Intelligibility and MIR features

With the set of 8 features composed of SRMR, MIR and P563 parameters, the SVR (C = 0.3, epsilon = 1 and gamma = 0.01) obtains a result of 0.3628 on the development set (Feat4). It is a low value in comparison with the previous configurations, but it is worth mentioning that this parameterization scheme uses a very low number of coefficients.

When this feature set is combined with the baseline acoustic features selected by Elastic Net after removing frames of low energy (Feat5), the Spearman correlation provided by SVR (C = 0.3, epsilon = 1 and gamma = 0.01) increases up to 0.5629.

4.6. Ranking SVM and Random Forests

Ranking SVM does not achieve improvements over SVR.

The use of a standard Random Forest software resulted in achieving 0.57 on the development set (after removing silence and applying Elastic Net), proving the ability of the Random Forest for improving the ranking results with a combination of bagging and regression trees. This result can only be slightly improved by optimising the Rand Forest configuration parameters: 0.587 (using 300 bags, the number of threshold candidates for tree spliting equal to 256, stop after 100 rounds without improvement, minimum number of samples per leaf equal to =1, unitary sub-sampling rate, feature sampling rate equal to 0.2 100 leafs per tree, 1 tree per bag and learning rate equal to 0.1).

Intelligibility can also be used in an ensemble of classifiers. Using the predicted values of this last and best classifier, we trained a second classifier (Weka's linear regression) to improve the Spearman correlation by means of the intelligibility features, achieving the best result (0.609), although for getting this result we have used 2-fold cross-validation on the development set.

5. Conclusions

In this paper we have tested many techniques for predicting the degree of Parkinsons condition using exclusively the patients voice, on the development set. In spite of the quality of the baseline provided by the Challenge organisers (0.492), we have been able to obtain several significant improvements on the evaluation metric of the challenge: the Spearman correlation.

First we have unsupervisedly removed low-energy frames, which cannot contribute to improve the prediction (0.535). Then, given the number of features in the baseline experiment is quite huge, we have applied Elastic Net for selecting the best subset of features, achieving a second improvement (0.557).

As the challenge is evaluated through the Spearman correlation, the use of classifiers specialized in ranking was a clear choice. Random Forests have obtained significant improvements (0.587).

Finally, an ensemble of classifiers obtained the best results, combining the output of the best Random Forest, with intelligibility features (0.609).

In order to deal with the mismatch between the training/development set and the test set, we have applied denoising techniques [26]. The best result obtained on the test set was 0.312.

6. Acknowledgements

The work leading to these results has been partly supported by Spanish Government grants TEC2014-53390-P and DPI2014-53525-C3-2-R, and from the European Union under grant agreement number 287678 (SIMPLE4ALL). Authors also thank all the other members of the Speech Technology Group at UPM and Grupo de Procesado Multimedia at UC3M for the continuous and fruitful discussion on these topics.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, 2015.

[2] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[3] S. Skodda, H. Rinsche, and U. Schlegel, "Progression of dysprosody in parkinson's disease over time–a longitudinal study," *Movement Disorders*, vol. 24, no. 5, pp. 716–722, 2009.

[4] A. Tsanas, M. A. Little, P. McSharry, and L. O. Ramig, "Enhanced classical dysphonia measures and sparse regression for telemonitoring of parkinson's disease progression," in *ICASSP 2010 – IEEE International Conference on Acoustics Speech and Signal Processing, Proceedings*, 2010, pp. 594–597.

[5] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[6] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease," in *LREC 2014 – Ninth International Conference on Language Resources and Evaluation, Proceedings*, 2014, pp. 342–347.

[7] A. Bayestehtashk, M. Asgari, S. I., and M. J., "Fully automated assessment of the severity of parkinson's disease from speech," *Computer, Speech & Language*, vol. 29, no. 1, pp. 172–185, 2015.

[8] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *ACM Multimedia (MM), Proceedings*, 2013, pp. 835–838.

[9] J. Foote, M. Cooper, and U. Nam, "Audio retrieval by rhythmic similarity," in *ISMIR 2002 – 3rd International Conference on Music Information Retrieval, Proceedings*, 2002.

[10] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*. Springer-Verlag, 1998.

[11] K. K. Jensen, *Timbre Models of Musical Sound: From the model of one sound to the model of one instrument*. DIKU, University of Copenhagen, 1999.

[12] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *DAFx 2007 – International Conference on Digital Audio Effects, Proceedings*, 2007.

[13] *Albayzin*. http://catalog.elra.info/product_info.php?products_id=746.

[14] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2009.

[15] ITU-T P.563, *Single-ended method for objective speech quality assessment in narrowband telephony applications*. Int. Telecom. Union, 2004.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[17] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Support Vector Regression Machines. Advances in Neural Information Processing Systems 9, NIPS 1996, pp. 155–161*. MIT Press, 1997.

[18] T. Joachims, "Training linear SVMs in linear time," in *KDD 2006 – ACM Conference on Knowledge Discovery and Data Mining, Proceedings*, 2006.

[19] *SVM-rank toolkit*. http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] V. Dang, *Ranklib - a library of ranking algorithms*. http://www.cs.umass.edu/ vdang/ranklib.html.

[22] R. Busa-Fekete, B. Kégl, T. Éltető, and G. Szarvas, "Learning-to-rank; multi-class classification; class probability calibration; regression based calibration; ensemble methods," *Machine Learning*, vol. 93, no. 2–3, pp. 261–292, 2013.

[23] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[24] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *CIKM 2009 – 18th ACM Conference on Information and Knowledge Management, Proceedings*, 2009, pp. 621–630.

[25] J. H. Friedman, *Greedy function approximation: A gradient boosting machine*. Stanford: Technical Report, IMS Reitz Lecture, 1999.

[26] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *ICASSP 2002 – IEEE International Conference on Acoustics Speech and Signal Processing, Proceedings*, 2002.