# UNOBTRUSIVE MONITORING OF SPEECH IMPAIRMENTS OF PARKINSON'S DISEASE PATIENTS THROUGH MOBILE DEVICES

T. Arias-Vergara[1], J.C. Vásquez-Correa[1,2], J.R. Orozco-Arroyave[1,2], P. Klumpp[2], and E. Nöth[2]

[1] Faculty of Engineering. Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
[2]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Corresponding author: tomas.arias@udea.edu.co

## ABSTRACT

Parkinson's disease (PD) produces several speech impairments in the patients. Automatic classification of PD patients is performed considering speech recordings collected in non-controlled acoustic conditions during normal phone calls in a unobtrusive way. A speech enhancement algorithm is applied to improve the quality of the signals. Two different classification approaches are considered: the classification of PD patients and healthy speakers and a multi-class experiment to classify patients in several stages of the disease. According to the results it is possible to classify PD patients and healthy controls with a AUC of up to $0.87$. This work is a step forward to the development of telemonitoring systems to assess the speech of the patients.

***Index Terms***— Parkinson's disease, speech impairments, mobile devices, speech enhancement, classification.

## 1. INTRODUCTION

Parkinson's disease (PD) is a neurological disorder characterized by the progressive loss of dopaminergic neurons in the mid-brain, producing several motor and non-motor impairments in the patients [1]. The motor symptoms include, among others, bradykinesia, rigidity, resting tremor, micrographia, and different speech impairments. The majority of PD patients develop several speech disorders [2], which may be considered as an early sign of further motor impairments [3]. Speech of PD patients is affected in several dimensions including phonation, articulation, and prosody [4]. Phonation impairments include inadequate closing of the vocal fold and vocal fold bowing [5]. The articulation problems are mainly related to reduced amplitude and velocity of lip, tongue, and jaw movements [6], while prosody refers to intonation, loudness, and rhythm during continuous speech.

There has been an interest in the research community to develop technology to monitor patients with neurodegenerative disorders using smartphones. For instance, a portable system for the automatic recognition of the syllables /pa-ta-ka/ is presented in [7]. The system consists of a tablet and a headset to capture the speech recordings from two group of speakers: patients with traumatic brain injuries and PD patients. The automatic recognition of /pa-ta-ka/ is performed in the mobile device using an automatic speech recognizer. Speech impairments are assessed using the syllable error rate. In [8], the authors develop an application to evaluate different PD symptoms related to dysphonia, postural instability, bradykinesia, and tremor. The assessment of PD symptoms is performed with a protocol to measure different motor impairments in voice, gait, dexterity, and balance. Although several motor impairments are considered, it is not clear whether the proposed system is suitable to assess each patient individually. On the other hand, there has been progress in methodologies to classify and monitor the PD symptoms using speech. In [9] the authors evaluated different phonation features to classify PD patients and healthy control (HC) speakers. They extract features from sustained vowels including stability and periodicity, noise measures, spectral wealth, and non-linear dynamics. Accuracies of up to $84\%$ were reported, depending on the analyzed vowel and on the feature set. In [10], the authors modeled different articulatory deficits in PD patients in the rapid repetition of the syllables /pa-ta-ka/, and reported an accuracy of $88\%$ discriminating between PD patients and HC speakers. Prosody features were computed in [11]. The authors consider voiced segments as speech unit to compute features based on the fundamental frequency $F_0$ contour, energy contour, duration, and pitch periods to classify PD patients and HC speakers, and to classify the patients according to their neurological state in a 3-class approach (low, middle, and severe) state. The authors report an accuracy of up to $74\%$ classifying PD patients and HC speakers, and of $37\%$ for the 3-class problem. Recently, in [12], the authors evaluated the effect of several feature extraction methods to classify PD patients and HC subjects in different acoustic conditions. They concluded that the effect of acoustic conditions is not critical when train and test sets are computed with recordings in the same scenario (matched); however, for the mismatched scenario the impact in the classification is higher. This paper introduces a methodology to evaluate the speech impairments of PD patients in recordings captured with smartphones during a normal phone call. The aim is to perform an unobtru-

sive monitoring of the patients through speech. We consider several approaches based on classical feature extraction to assess phonation, articulation, and prosody. Additionally, we consider a deep learning model based on convolutional neural networks (CNNs) to assess articulation impairments. A Speech Enhancement (SE) algorithm is also applied to the phone calls to improve the quality of the recordings. According to the results, prosody is the most suitable speech dimension to evaluate speech impairments of PD patients during spontaneous speech. To the best of our knowledge this is the first contribution considering a mismatched scenario evaluating speech impairments of PD patients under real phone conversations.

## 2. METHODS

### 2.1. Speech enhancement

We consider the log-minimum mean square error estimator (logMMSE) algorithm introduced in [13] to improve the quality of the phone calls. The method finds an estimator of the noise-free speech signal $x(t)$ that minimizes the mean square error between the log-spectrum of the noise-free speech signal and its estimator. We find an estimator for the enhanced signal directly from the amplitude spectrum of the noisy observable signal $y(t)$, multiplied by a non-linear gain function that depends only on the a priori signal to noise ratio. The gain function is estimated with the first 120 ms of the noisy speech signal and updated in each silence part.

### 2.2. Feature extraction and classification

Three different feature sets are computed based on phonation, articulation, and prosody analysis. Phonation features are extracted from voiced segments to model the temporal and amplitude variation of the vocal fold vibration. Articulation impairments are modeled considering spectral measures and the energy content of the onset/offset transitions. Prosodic impairments are modeled considering the contours of the fundamental frequency $F_0$ and the energy. The code to extract the features is freely available[1] and the details about the computation of these feature sets can be found in [4] and [14].

**Phonation–** The phonation analysis in continuous speech is performed by extracting voiced segments from the utterance. The feature set includes seven descriptors such as jitter and shimmer, the first and second derivatives of $F_0$, long term perturbation features such as the amplitude perturbation quotient and the pitch perturbation quotient and the energy. The mean, standard deviation, skewness, and kurtosis are computed from the descriptors, forming a 28-dimensional feature vector per utterance.

**Articulation–** The articulatory capability of the patients is evaluated with information from the onset/offset transitions to model the difficulties of patients to start/stop the movement of the vocal folds. The set of features extracted from the onset and offset include 12 Mel-Frequency Cepstral Coefficients (MFCCs) with their first and second derivatives, and the log energy of the signal distributed into 22 Bark bands. The first and second formant frequencies and their derivatives are also considered to assess the articulation deficits of the patients. The total number of descriptors corresponds to 87. Four functionals are also computed, obtaining a 488-dimensional feature-vector per utterance.

**Prosody–** The prosody features are based on duration, the $F_0$ contour and the energy contour. We compute 13 features per utterance including the average, standard deviation, and maximum value of $F_0$; the variability of $F_0$ expressed in semitones; the average, standard deviation, and maximum value of the energy contour; the voiced rate, the average and standard deviation of the duration of voiced segments, the pause rate, and the average and standard deviation of the duration of pauses.

**Classification–** The automatic classification of PD patients and HC subjects is performed with a Support Vector Machine (SVM) with margin parameter $C$ and a Gaussian kernel with parameter $\gamma$. The values of $C$ and $\gamma$ are optimized through a grid-search into powers of ten with $10^{-4} < C < 10^4$ and $10^{-6} < \gamma < 10^3$. The selection criterion is based on the performance obtained in the training stage. Due to the low number of speakers, the SVM is tested following a Leave-One-Out Cross-Validation strategy.

### 2.3. CNN modeling

The onset and offset transitions are detected similar to the previous articulation features. The transition is detected, and 80 ms of the signal are taken to the left and to the right of each border, forming "chunks" of signals with 160 ms length. Each chunk is transformed into a time-frequency representation using the short-time Fourier transform (STFT) and used as input to a CNN [15]. The CNN extracts the most suitable features from the STFT and makes the final decision about whether the utterance corresponds to a PD patient or a HC speaker, or classify the speaker according to the level of the speech item in the third part of the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS-III). Rectifier linear activation functions are used, and dropout is included in the training stage to avoid over–fitting [16]. The architecture of the CNN implemented in this study is summarized in Figure 1. It consists of four convolutional layers, two max-pooling layers, and two fully connected hidden layers followed by the output layer to make the final decision using a sigmoid activation function. The CNN is trained using the stochastic gradient descent algorithm. The cross–entropy between the training labels $y$ and the model predictions $\widehat{y}$ is used as the loss function. In addition, the root mean square propagation is considered as a mechanism to adapt the learn-

[1] https://github.com/jcvasquezc/DisVoice

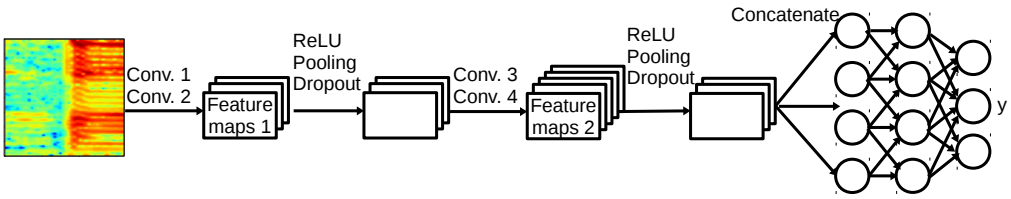


**Fig. 1**. Architecture of the CNN implemented in this study

ing rate in each iteration. The method divides the learning rate $\eta$ by an exponentially decaying average of squared gradients [17]. CNN's hyper-parameters are optimized following a Bayesian optimization approach [18]. We optimize the kernel size for the convolutional layers (from 3 to 7 ), the number of feature maps (from 8 to 64), the size of the fully connected layers (from 32 to 256), the initial learning rate (from 0.001 to 0.01), and the probability of dropout (from 0.1 to 0.7) between consecutive layers.

## 3. DATA

### 3.1. Train data

An extended version of the PC–GITA database [19] is considered to train the models. The data contain speech utterances of 68 PD patients and 50 HC subjects balanced in age and gender. All of them are Colombian Spanish native speakers. The HC speakers were recorded once, while 33 of the patients were recorded in several sessions between 2012 and 2016. Most of the patients were recorded in two or three sessions. The speech signals were recorded with a sampling frequency of 44.1 kHz and 16-bit resolution. The speech recordings were re-sampled to 16 kHz to meet the sampling frequency of the test set. The participants pronounce a monologue according to their daily activities, with an average duration of 79.1±43.8 seconds. Additional information about the train data is shown in Table 1. In addition, the distributions of the MDS-UPDRS-III score and the speech item of the same scale are shown in Figure 2. We define three classes according to the histogram of the speech item: zero for low level speech impairments, one for middle stage speech impairments, and greater than one for severe speech deficits.

**Table 1**. Meta-information of the training set. $\mu$: average, $\sigma$: standard deviation.

| | PD patients | | HC subjects | |
|---|---|---|---|---|
| | male | female | male | female |
| Number of subjects | 35 | 33 | 25 | 25 |
| Age ($\mu \pm \sigma$) | $61.8 \pm 10.5$ | $60.1 \pm 8.1$ | $60.5 \pm 11.6$ | $61.4 \pm 7.0$ |
| Age range | 33–81 | 42–75 | 31–86 | 49–76 |
| Disease duration ($\mu \pm \sigma$) | $8.5 \pm 5.2$ | $13.5 \pm 11.8$ | | |
| Range of disease duration | 1–20 | 1–43 | | |
| MDS-UPDRS-III ($\mu \pm \sigma$) | $43.0 \pm 22.7$ | $37.7 \pm 13.3$ | | |
| Range of MDS-UPDRS-III | 6–93 | 19–71 | | |

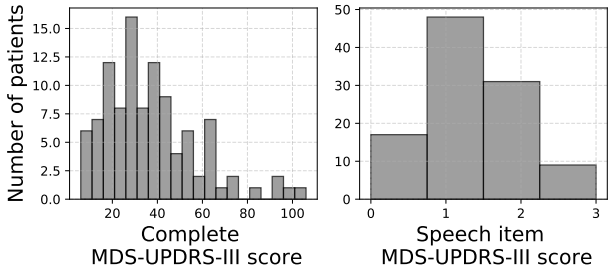


**Fig. 2**. Histograms for the complete MDS-UPDRS-III score (left) and the item related to speech (right).

### 3.2. Test data

The speech of 17 PD patients (9 male, 8 female) was recorded using the *Apkinson* mobile application [20]. The participants were asked to make a phone call and sustain an spontaneous conversation. Several speech aspects were evaluated based on such conversations. None of the speakers in the test set were included in the train set. The average duration of the recordings is 62.9±49.9 seconds. The recordings were captured using different smartphones and in different acoustic conditions with a sampling frequency of 16 kHz. The test data contain also 7 HC subjects whose age ranges from 51 to 79 years.

**Table 2**. Meta-information of the test set. $\mu$: average, $\sigma$: standard deviation.

| | PD patients | | HC speakers | |
|---|---|---|---|---|
| | male | female | male | female |
| Number of subjects | 10 | 7 | 2 | 5 |
| Age ($\mu \pm \sigma$) | $60.9 \pm 8.2$ | $64.5 \pm 8.6$ | | |
| Range of age | 53-80 | 56-83 | 51-79 | |
| Range of MDS-UPDRS-III | 17-69 | 24-41 | | |

## 4. EXPERIMENTS AND RESULTS

Two experiments are performed: (1) to classify PD patients and HC subjects, and (2) to classify the patients into the three stages defined above according to the speech item of the MDS-UPDRS-III. The classifiers are trained with the speakers from Table 1 and tested with the speakers from Table 2, which were recorded using *Apkinson*. Speech recordings

were captured in a mismatched scenario, i.e., the test set considers mobile phone recordings while the train set is formed with utterances recorded in controlled acoustic conditions.

### 4.1. PD detection

The SVM and CNN were tested using the original phone calls, i.e., without the SE algorithm and the phone calls processed with SE. The results are presented in Table 3. For the case of the SVM, the accuracies range from $58\%$ to $75\%$ in the original phone calls. The results improve in up to $21\%$ (absolute) when the recordings are processed with the SE algorithm. Such improvement can be observed clearly with the AUC obtained for each experiment. In addition, the highest improvement is obtained for prosody features, where the AUC improves from $0.59$ to $0.87$.

**Table 3**. Results for classification of PD patietns and HC subjects. **ACC (%)**: Accuracy. **SEN (%)**: Sensitivity. **SPE (%)**: Specificity. **AUC**: Area under the ROC curve. **Fusion**: Combination of phonation, articulation, and prosody features.

| **Features** | **Original** | | | | **Speech enhancement** | | | |
|---|---|---|---|---|---|---|---|---|
| | **ACC** | **SEN** | **SPE** | **AUC** | **ACC** | **SEN** | **SPE** | **AUC** |
| Phonation | 66 | 76 | 42 | 0.59 | 71 | 75 | 50 | 0.66 |
| Articulation | 58 | 68 | 20 | 0.61 | 71 | 81 | 50 | 0.61 |
| Prosody | 58 | 70 | 28 | 0.59 | 79 | 88 | 63 | 0.87 |
| Fusion | 66 | 88 | 14 | 0.66 | 62 | 75 | 32 | 0.62 |
| CNN | 61 | 82 | 0 | 0.53 | 58 | 76 | 14 | 0.54 |

For the CNN, only the articulatory capability of the patients was evaluated, as in previous studies [15]. The results show that the performance of the CNN was slightly lower than for the SVM after the SE. Although, previous experiments have shown the suitability of the CNN to model the articulation impairments of the patients, it seems like the CNN is not able yet to adapt to the different acoustic conditions on the test set given mismatched channel. This could be explained due to the limited size of the test set.

### 4.2. Multi-class experiment

The automatic multi-class assessment is performed according to the speech item of the MDS-UPDRS-III score (3.1) using a multi-class SVM following a one vs all strategy. Class 0 includes speakers with a score of 0 (HC in the test set), class 1 includes patients with a score of 1, and class 2 includes speakers with scores higher than 2. The optimal parameters ($C = 1; \gamma = 10^{-2}$) were found in an internal cross-validation in the train set. Table 4 shows the results obtained when the prosody features are considered. For the train set, it can be observed that most of the speakers with a score of 0 (Class 0) and higher than 2 (Class 2) are assigned to the correct class. For Class 1, most of the classification errors occurs in Class 0. This result can be explained considering that a score of 1 is assigned to patients with minimal speech problems. For the test set, all of the speakers from Class 0 are classified correctly, which indicates that the system is capable of identify healthy speech. None of the patients from Class 1 were assigned to their true class, which may be explained due to the mismatched in the acoustic conditions, or due to the class unbalance in the train set.

**Table 4**. Confusion matrix obtained with the prosody features when SE is applied. Results in $\%$

| | Train set Predicted class | | | Test set Predicted class | | |
|---|---|---|---|---|---|---|
| Target class | Class 0 | Class 1 | Class 2 | Class 0 | Class 1 | Class 2 |
| Class 0 | 96 | 2 | 2 | 100 | 0 | 0 |
| Class 1 | 53 | 38 | 9 | 50 | 0 | 50 |
| Class 2 | 29 | 4 | 67 | 46 | 0 | 54 |

Train set: Cohen's $\kappa$: $0.56$; Accuracy: $72\%$; UAR: $67\%$
Test set: Cohen's $\kappa$: $0.31$; Accuracy: $58\%$; UAR: $51\%$

## 5. CONCLUSIONS

A method for the unobtrusive monitoring of PD patients from speech is presented in this paper. The speech of the patients is captured during a phone call using smartphones during spontaneous conversations. Several feature sets are computed to assess the phonation, articulation and prosody impairments. In addition, an SE algorithm is applied to improve the quality of the recordings. We evaluate the suitability of the proposed methodology with two experiments: the classification of PD patients and HC speakers and a multi-class classification according to the speech item of the MDS-UPDRS-III scale. The SE algorithm is suitable to improve the quality of the speech recordings from the mobile phone and also the results of the classification of patients given the train/test channel mismatched acoustic conditions. Additionally, The best results were obtained when the prosody features are considered. This result confirms that the variations of the speech during a free conversation, which are intended to be assessed with the phone calls are suitable to assess the speech deficits of PD patients. Further studies may be performed with more data from normal, spontaneous and unobtrusive conversations. Data collection using *Apkinson* is still ongoing, thus in the near future we expect to perform more experiments for further development of the mobile application.

# Acknowledgments

This work was financed by CODI from University of Antioquia by the grant Number PRV16-2-01 and 2015–7683. This work also contributes to the research project DysarTrain which aims to provide an automatic therapy tool for patients suffering from dysarthric speech. T. Arias-Vergara and J. C. Vásquez-Correa acknowledge to the Training Network on Automatic Processing of PAthological Speech (TAPAS) funded by the Horizon 2020 programme of the European Commission

## 6. REFERENCES


[1] O. Hornykiewicz, "Biochemical aspects of Parkinson's disease," *Neurology*, vol. 51, no. 2 Suppl 2, pp. S2–S9, 1998.

[2] J. A. Logemann et al., "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[3] J. Hlavnicka et al., "Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Nature Scientific Reports*, vol. 7, no. 12, pp. 1–13, 2017.

[4] J.R. Orozco-Arroyave, *Analysis of speech of people with Parkinson's disease*, Logos-Verlag, Berlin, Germany, 1st edition, 2016.

[5] David G Hanson et al., "Cinegraphic observations of laryngeal function in Parkinson's disease," *The Laryngoscope*, vol. 94, no. 3, pp. 348–353, 1984.

[6] H. Ackermann and W. Ziegler, "Articulatory deficits in parkinsonian dysarthria: an acoustic analysis.," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 54, no. 12, pp. 1093–1098, 1991.

[7] F. Tao et al., "A portable automatic PA-TA-KA syllable detection system to derive biomarkers for neurological disorders," in *Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association*, 2016, pp. 362–366.

[8] A. Zhan et al., "High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection," *arXiv preprint arXiv:1601.00960*, 2016.

[9] J. R. Orozco-Arroyave et al., "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1820–1828, 2015.

[10] M. Novotnỳ et al., "Automatic evaluation of articulatory disorders in Parkinson's disease," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 22, no. 9, pp. 1366–1378, 2014.

[11] T. Bocklet et al., "Automatic Evaluation of Parkinson's Speech – Acoustic, Prosodic and Voice Related Cues," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, pp. 1149–1153.

[12] J. C. Vásquez-Correa et al., "Effect of acoustic conditions on algorithms to detect parkinson's disease from speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5065–5069.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[14] J. R Orozco-Arroyave et al., "Neurospeech: An open-source software for Parkinson's speech analysis," *Digital Signal Processing (In press)*, 2017.

[15] J. C. Vásquez-Correa et al., "Convolutional neural network to model articulation impairments in patients with parkinson's disease," in *18th Annual Conference of the Speech and Communication Association (INTERSPEECH)*, 2017, pp. 1–5.

[16] N. Srivastava et al., "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[18] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 2951–2959.

[19] J. R. Orozco-Arroyave et al., "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Language Resources and Evaluation Conference, (LREC)*, 2014, pp. 342–347.

[20] P. Klumpp et al., "Apkinson–A Mobile Monitoring Solution for Parkinson's Disease," in *18th Annual Conference of the Speech and Communication Association (INTERSPEECH)*, 2017, pp. 1839–1843.