END-2-END MODELING OF SPEECH AND GAIT FROM PATIENTS WITH PARKINSON'S DISEASE: COMPARISON BETWEEN HIGH QUALITY VS. SMARTPHONE DATA

J. C. Vasquez-Correa[1,2⋆] *T. Arias-Vergara*[1,2,3] *P. Klumpp*[1] *P. A. Perez-Toro*[1,2]
J. R. Orozco-Arroyave[1,2] *E. Nöth*[1]
[1]Pattern Recognition Lab. Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany
[2] Faculty of Engineering. Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
[3] Department of Otorhinolaryngology. Ludwig-Maximilians Universität, München, Germany
⋆corresponding author: juan.vasquez@fau.de

ABSTRACT

Parkinson's disease is a neurodegenerative disorder characterized by the presence of different motor impairments. Speech and gait signals have been analyzed to detect the presence of the disease and the severity in patients. However, most studies have been performed in controlled conditions using high quality data, which make those studies not suitable for a continuous at-home evaluation of the state of the patients. The developed technology should be evaluated in more realistic scenarios, for instance using smartphone data. We propose the use of state-of-the-art deep learning techniques to evaluate the speech and gait symptoms of patients. The proposed methods are evaluated in two scenarios to cover both high quality and smartphone data. The results indicate that it is possible to classify patients and healthy subjects with accuracies over 92% in both scenarios. The proposed methods are also promising to evaluate the severity of the speech symptoms and the global motor state of the patients.

***Index Terms*—** Parkinson's disease, Deep learning, Gait analysis, Speech analysis, Smartphones.

1. INTRODUCTION

Parkinson's disease (PD) is a neuro-degenerative disorder that produces different motor symptoms in the patients, including tremor, rigidity, and bradykinesia, among others [1]. 70-90% of patients develop a a speech impairment called hypokinetic dysarthria [2], which manifests in the imprecise articulation of consonants, monoloudness, and monopitch, among other symptoms. The traditional assessment of the disease depends on the experience of the clinician performing the screening, which makes the diagnosis of disease as well as its degree of severity difficult. It is important to identify the earliest symptoms of PD in order to be able to treat the disease in the prodromal phase, and to evaluate how severe the symptoms of a patient are in order to prescribe a better treatment.

Several studies have modeled the speech of PD patients in terms of phonation, articulation, prosody, and intelligibility [3, 4]. These traditional methods are based on the computation of hand-crafted features such as jitter, shimmer, or formant frequencies that may not completely model all the phenomena that appear due to the presence of the disease and the dysarthria level of patients. There are recent studies that have proposed the use of deep learning methods to model the speech of PD patients [5, 6, 7]. Most of them consider convolutional neural networks (CNN) to process time-frequency representations of the speech signals like Mel-spectrograms. The authors usually focus only in the classification of PD vs. healthy control (HC) subjects, leaving aside the evaluation of the disease severity. The accuracy reported in those studies ranges from 80% to 90% for the classification of PD vs. HC subjects. The research community has shown also a growing interest in the automatic gait analysis of PD. The assessment is performed commonly with inertial sensors attached to the body of the patients [8, 9] and with force-sensitive sensors placed inside the shoes of the participants [10]. By using inertial sensors, it is possible to detect and to characterize specific movements and to monitor activities of daily living of PD patients [11]. Most of the studies have considered kinematic features based on the duration and velocity of the steps [10, 12, 13]. Other studies have considered spectral features to evaluate the harmonic structure of the gait process [14, 15], or non-linear dynamics methods to model long-range autocorrelations and stability patterns of the walking process [16, 17, 18]. There are few studies that have considered deep learning models to evaluate the gait of PD patients using the raw gait signals in order to the neural network automatically learns the most appropriate features [19, 20].

Most studies to model speech and gait symptoms of PD patients have been performed in controlled conditions, using high quality speech data, and with external inertial sensors attached to the body of the participants [21]. These aspects make many of the proposed studies available for the clinical practice but not for an at-home evaluation of the state of the patients. The technology to monitor the state of PD patients should be evaluated in more realistic scenarios, for instance using smartphones. A more reliable assessment of the patients at-home can be performed using the microphone and the inertial sensors available in smartphones, which can be used to evaluate different motor impairments in the speech production, and in the upper and lower limbs.

We propose the use of state-of-the-art deep learning techniques to evaluate the speech and gait symptoms of PD pa-

tients. The proposed methods are evaluated in two scenarios to cover both high quality data, which is normally captured in a clinical evaluation, and smartphone data, which can be used to monitor the state of the patients at-home. In both scenarios, the methods are used to classify PD vs. HC subjects, and to evaluate the severity of the motor symptoms. The results indicate that it is possible to classify PD patients and HC subjects with accuracies over 92% using both high quality and smartphone speech data, and with accuracies over 94% using gait data, in both scenarios. We believe that within the next decade, monitoring of motor symptoms of PD patients will gradually shift from the clinic to at-home, where a continuous monitoring can be performed. The next step will be the application of the proposed methods in the longitudinal and individual evaluation of the symptoms of the patients, in order to monitor the progression of the disease per patient.

2. MATERIALS

2.1. Multimodal corpus

The data include high quality speech and gait signals from 106 PD patients and 105 HC subjects, Colombian Spanish native speakers. These data are age- and gender-balanced. 94 of the patients were labeled according to the third section of the movement disorder society - unified Parkinson disease rating scale (MDS-UPDRS-III). Additionally, the speech recordings from 93 of the PD patients and from 48 of the HC subjects were labeled according to the modified Frenchay dysarthria assesment (m-FDA) scale, which evaluates the dysarthria severity of the participants [22].

The speech protocol includes the utterance of six diadochokinetic (DDK) exercises, the reading of 10 sentences, a read text with 36 words phonetically balanced, and a monologue where the participants were asked to speak about their daily routine. The speech signals were recorded with a sampling frequency of 16 kHz and 16-bit resolution. The gait signals were captured with the eGaIT system, which consists of a 3D-accelerometer (range $\pm$6g) and a 3D gyroscope (range $\pm$500°/s) attached to the external side (at the ankle level) of the shoes [9]. Data from both feet were captured with a sampling frequency of 100 Hz and 12-bit resolution. The exercises included 20 meters walking with a stop after 10 meters (2x10), 40 meters walking with a stop every 10 meters (4x10), 20 meters walking with stops every three meters (Stop & go), *heel-toe tapping*, and the *time up and go* (TUG) test.

2.2. Apkinson corpus

This corpus was collected using the Apkinson android application [23], which was designed to record several signals using the microphone and accelerometer available on smartphones. The data contain speech and movement signals collected from 38 PD patients and 60 HC subjects. 26 of the patients were labeled with the MDS-UPDRS-III. None of the participants in the HC group presented any neurological or movement disorder. The age and gender distributions per class is also balanced for the Apkinson corpus. The speech tasks include the same six DDK exercises from the multimodal corpus, the reading of the 10 sentences, and a monologue based on the description of the cookie theft picture from the Boston diagnostic aphasia examination. The movement signals contain 7 tasks captured with the inertial sensors of the smartphone, and include: (1) *Posture*, where the patient stands up straight during 30 seconds, (2) *circles*, where the patient has to make circles with the extended arm, (3) *pronation/supination*, where the patient stretches out the arm with the downward palm, and then turn the palm up-down, several times, (4) *finger to nose*, where the patient extends the arm and then touches his/her nose and extends the arm again, several times, (5) *postural tremor*, where the patient extends the arm and holds the smartphone in this position for at least 10 seconds, (6) *4x10*, where patients perform a short path walking four times, and (7) *Free Gait*, where patients perform a normal walk exercise during two minutes. For the case of the walking exercises we ask the patients to put the smartphone in their pockets. For the case of the hand movement exercises the patients take the smartphone with their hands.

3. METHODS

3.1. Speech modeling

The proposed model to process speech signals is based on CNNs using Mel-spectrograms as input. We computed the Mel-spectrum for windows of 32ms length and a time-shift of 4ms. This Mel spectrum is computed with a frequency resolution of 512 points and 64 Mel filters. We then stack together 126 of these Mel spectra to form a Mel spectrogram with 500ms length, which is used as input for our proposed CNN. These parameters lead us to a time frequency representation of 126 time steps and 64 frequency bins. The spectrograms are modeled with a ResNet18 architecture, which has three residual blocks and 18 convolutional layers (see Figure 1). The skip connections help to control the vanishing gradient problem when we have deeper models. Dropout layers were considered to regularize the output of the residual blocks. The final decision is made by a fully connected layer with a Softmax activation function.

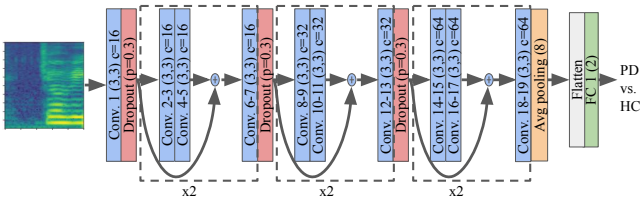

Fig. 1. ResNet18 model to process the Mel-spectrograms of the speech signals. **FC**: Fully connected layers. **c**: number of output channels. Values in parenthesis indicate the size of the conv. filters and the number of neurons in the FC layers.

3.2. Gait & Movement modeling

We propose a deep learning model based on 1D-convolutions to process the raw gait signals. Figure 2 illustrates the proposed architecture to model the gait signals of the patients. The input corresponds to 3 seconds-length frames of the gait signals. For the case of the multimodal corpus, the input is formed with 12 channels corresponding to the 3D-accelerometer and 3D-gyroscope attached to the left and right foot. The input for the Apkinson data includes only three channels from the 3D-accelerometer from the smartphone. The duration was chosen to guarantee at least 3 periods of the gait signals. The input then passes through a set of two 1D-convolutional layers, which learn a filter-bank. The filtered signals then pass through a stack of two bidirectional gated recurrent unit (GRU) layers to model the temporal structure of the sequences. The last part of the network is an attention mechanism, which assigns more weights to specific parts of the gait sequence, such as pauses, the swing phase, the stance phase, or the beginning/stopping of the gait task.

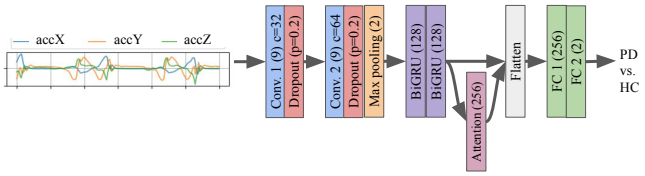

Fig. 2. Deep learning model for end-to-end gait modeling of PD patients. **FC**: Fully connected layers. **c**: number of output channels. Values in parenthesis indicate the size of the conv. filters and the number of neurons in the FC layers.

4. EXPERIMENTS AND RESULTS

Different experiments are performed to classify PD patients vs. HC subjects and to evaluate the disease severity of the participants. All models are validated with a 10-fold stratified cross-validation strategy. The first experiment corresponds to the classification of PD patients vs. HC subjects using speech signals from the multimodal and Apkinson corpora. The results are shown in Table 1. The accuracy for the multimodal corpus range from 88.8% to 92.4%, similar to the one obtained for the Apkinson data, which ranges from 86.7% to 92.2% depending on the speech task. These results confirm those reported previously about data collected with smartphones having enough quality to classify speech signals from PD patients [24, 25, 26]. This study is the first one to confirm that similar results are obtained both with high quality and smartphone data using a full deep learning approach. Results reported here also consider higher amounts of data than the studies previously reported.

The results classifying PD patients and HC subjects using the gait & movement signals are shown in Table 2. The accuracy for the multimodal corpus ranges from 90.6% to 98.7%. The highest accuracy is observed in the *Stop & Go* task, which is the one when the patients have to perform more start/stop movements of the lower limbs, causing Freezing of Gait (FoG) episodes in the patients that are modeled with our proposed approach. The results observed for the Apkinson corpus indicate that gait exercises like *4x10* and *Free gait* produce the highest accuracies, and the results are similar to the ones obtained with the high quality inertial sensors used in the multimodal corpus. Hand movement tasks like the *finger to nose* and the *circles* produce moderate accuracies. Conversely, tasks such as *postural tremor*, *posture*, or *pronation/supination* are not accurate for the classification using the proposed model. These particular exercises have a very low dynamic compared with the walking tests. The lack of accuracy for these particular tasks can be explained because such small temporal variability is not properly captured with the smartphone sensors. Unfortunately, we do not have data collected with the high-quality sensors to address these tasks and validate these results. Other methods can be proposed to model the information produced by these types of tasks.

Table 1. Classification of PD vs. HC subjects using speech signals. Results in terms of average (standard deviation).

Task	UAR [%]	SENS [%]	SPEC [%]	AUC
Speech Multimodal corpus				
DDK	88.8 (2.4)	82.0 (2.3)	95.5 (2.9)	0.949
Sentences	87.5 (1.0)	83.3 (2.3)	91.7 (2.6)	0.949
Read text	92.4 (6.4)	87.0 (6.5)	97.7 (9.9)	0.974
Monologue	88.8 (6.5)	88.2 (6.5)	89.4 (9.9)	0.948
Speech Apkinson corpus				
DDK	92.2 (2.6)	96.1 (1.4)	88.2 (4.6)	0.932
Sentences	86.7 (1.9)	94.4 (2.9)	79.0 (2.6)	0.926
Monologue	87.2 (4.0)	87.9 (4.4)	86.5 (8.2)	0.910

UAR: unweighted average recall, **SENS**: sensitivity, **SPEC**: specificity, **AUC**: Area under the ROC curve.

Table 2. Classification of PD vs. HC subjects using gait signals. Results in terms of average (standard deviation).

Task	UAR [%]	SENS [%]	SPEC [%]	AUC
Gait Multimodal corpus				
2x10	96.6 (2.4)	93.2 (3.4)	100.0 (1.9)	0.998
4x10	96.5 (2.4)	94.9 (3.4)	98.1 (1.9)	0.997
Stop & Go	98.7 (2.4)	97.4 (3.4)	100.0 (1.9)	0.999
Heel Toe Tapping	90.6 (2.4)	88.8 (3.4)	92.3 (1.9)	0.963
TUG	96.5 (2.4)	94.8 (3.4)	98.1 (1.9)	0.988
Gait Apkinson corpus				
4x10	94.1 (5.2)	93.1 (9.0)	95.0 (4.6)	0.947
Free Gait	92.0 (5.2)	91.2 (9.0)	92.9 (4.6)	0.940
Finger to Nose	83.6 (6.3)	91.9 (11.4)	75.3 (1.3)	0.893
Circles	76.5 (4.6)	82.2 (10.2)	70.7 (1.0)	0.863
Postural Tremor	56.0 (7.5)	23.9 (16.1)	88.1 (1.1)	0.579
Posture	59.6 (5.2)	40.6 (9.0)	78.6 (4.6)	0.621
Pronation/Supination	51.9 (2.6)	5.0 (7.0)	98.8 (1.8)	0.602

UAR: unweighted average recall, **SENS**: sensitivity, **SPEC**: specificity, **AUC**: Area under the ROC curve.

For the third experiment we grouped the subjects from the multimodal corpus into three classes according to their dysarthria severity based on the m-FDA scale [22]. Unfortunately, we do not have labels of the m-FDA score for the subjects in the Apkinson data. The number of subjects per class was determined to guarantee balanced groups. In this

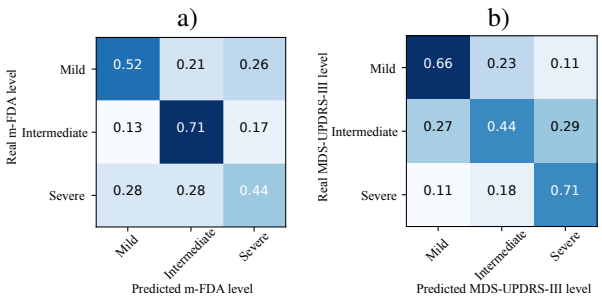

Fig. 3. Confusion matrices for the best results to evaluate: **a)** the dysarthria level of patients based on the m-FDA score, **b)** the motor state severity based on the MDS-UPDRS-III score.

experiment we classify subjects with mild, intermediate, or severe speech impairments, based on the values of their m-FDA scale. The division of the three groups was based on the 33th and 66th percentiles of the total m-FDA score for the subjects. The results are observed in Table 3. The highest accuracy was observed in the monologue task (55.7%). The confusion matrix for the best result is observed in Figure 3a). The class with the highest accuracy corresponds to the patients with intermediate dysarthria level, followed by patients in mild and severe states, respectively.

Table 3. Classification of the dysarthria severity using speech signals. Results in terms of average (standard deviation).

Task	Fscore	UAR [%]
Speech Multimodal corpus		
DDK	0.533 (0.02)	53.3 (2.2)
Sentences	0.515 (0.03)	51.9 (2.9)
Read text	0.516 (0.07)	52.1 (5.5)
Monologue	0.554 (0.07)	55.7 (5.5)

UAR: unweighted average recall

Finally, we classify the patients in different groups according to their motor severity based on the MDS-UPDRS-III. The patients were grouped into three classes according to their MDS-UPDRS-III score using the 33th and 66th percentiles of our data as a border between the three groups. The subjects in each group were labeled as patients in mild, intermediate, and severe states. The models were then trained to classify these three classes. The results are observed in Table 4. For the multimodal corpus, the highest accuracies are obtained with the *TUG* and with the *Stop & Go* tasks, similar to the results observed in the bi-class problem in Table 2. These results confirm the importance of such exercises for the assessment of the gait impairments of PD patients. The confusion matrix for the best result (TUG test) is observed in Figure 3b). The class with the highest accuracy corresponds to the patients in severe state, followed by patients in mild and intermediate states, respectively. Note also that the miss-classified patients from the mild and severe classes are mainly miss-classified as patients in intermediate state of the disease rather than in the other extreme class. Regarding the Apkinson corpus, the highest results are again obtained with the gait exercises (4x10 and Free Gait). For this case there are differences of up to 12% between the results obtained in the multimodal and Apkinson corpora. We believe that these differences are because of the reduced size of the Apkinson data to train the models for this particular and more difficult problem. Additional data using the Apkinson app should be collected and labeled to improve the results.

Table 4. Classification of the motor severity of patients using gait signals. Results in terms of average (standard deviation).

Task	Fscore	UAR [%]
Gait Multimodal corpus		
2x10	0.597 (0.125)	60.8 (9.6)
4x10	0.575 (0.125)	58.1 (9.6)
Stop & Go	0.612 (0.125)	62.9 (9.6)
Heel Toe Tapping	0.454 (0.125)	46.4 (9.6)
TUG	0.632 (0.125)	64.9 (9.6)
Gait Apkinson corpus		
4x10	0.467 (0.066)	49.2 (8.3)
Free Gait	0.485 (0.066)	52.9 (8.4)
Finger to Nose	0.334 (0.152)	38.6 (5.7)
Circles	0.301 (0.113)	28.5 (10.7)
Postural Tremor	0.363 (0.146)	45.6 (25.8)
Posture	0.392 (0.067)	38.8 (8.4)
Pronation / Supination	0.262 (0.125)	30.4 (16.0)

UAR: unweighted average recall

5. CONCLUSION

The present study proposes the use of deep learning methods to classify PD patients and HC subjects, and to evaluate the disease severity of the patients, using information from speech and gait signals. We evaluate the impact of the proposed approach in signals collected with smartphone sensors. The results show that it is possible to classify PD patients and HC subjects with accuracies of up to 92% using speech signals and of up to 98.7% using gait signals. In addition, the results indicate that there is not a visible difference in the accuracies observed when considering high quality vs. smartphone data. The disease severity of the patients is estimated with accuracies up to 55.7% for the speech impairments, and up to 64.9% for the global motor deficits. Additional data from smartphones should be collected and labeled to improve the results of the disease severity assessment. The next step will be to evaluate the proposed methods in the individual monitoring of the symptoms of the patients. In addition, we are currently running experiments combining speech and movement and the results look promising. We hope to include those results in future studies.

6. ACKNOWLEDGMENTS

This project received funding from the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287. This study was inspired by our work in the 2016 Jelinek Memorial Summer Workshop (JSALT), which was supported by JHU. Thanks also to CODI from Universidad de Antioquia grant No. PRG2017-15530.

7. REFERENCES

[1] O. Hornykiewicz, "Biochemical aspects of Parkinson's disease," *Neurology*, vol. 51, no. 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," *J. Speech Hear. Disord.*, vol. 43, no. 1, pp. 47–57, 1978.

[3] J. Rusz et al., "Imprecise vowel articulation as a potential early marker of parkinson's disease: Effect of speaking task," *J Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2171–2181, 2013.

[4] J. R. Orozco-Arroyave et al., "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digital Signal Processing*, vol. 77, pp. 207–221, 2018.

[5] J. C. Vásquez-Correa et al., "Convolutional neural network to model articulation impairments in patients with Parkinson's disease," in *INTERSPEECH*, 2017, pp. 314–318.

[6] J. Correia et al., "In-the-wild end-to-end detection of speech affecting diseases," in *ASRU*, 2019, pp. 734–741.

[7] M. Wodzinski et al., "Deep learning approach to parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification," in *EMBC*, 2019, pp. 717–720.

[8] Q. W. Oung et al., "Technologies for assessment of motor disorders in Parkinson's disease: a review," *Sensors*, vol. 15, no. 9, pp. 21710–21745, 2015.

[9] J. Barth, *Development and Validation of a Mobile Gait Analysis System Providing Clinically Relevant Target Parameters in Parkinson's Disease*, Logos-Verlag, Berlin, Germany, 1st edition, 2017.

[10] P. Ren et al., "Analysis of gait rhythm fluctuations for neurodegenerative diseases by phase synchronization and conditional entropy," *IEEE T. Neur. Sys. Reh.*, vol. 24, no. 2, pp. 291–299, 2016.

[11] L. Brognara et al., "Assessing gait in parkinson's disease using wearable motion sensors: a systematic review," *Diseases*, vol. 7, no. 1, pp. 18, 2019.

[12] M. Djurić-Jovičić et al., "Selection of gait parameters for differential diagnostics of patients with de novo Parkinson's disease," *Neurological research*, vol. 39, no. 10, pp. 853–861, 2017.

[13] C. Caramia et al., "IMU-based classification of Parkinson's disease from gait: A sensitivity analysis on sensor location and feature selection," *IEEE J. Biomed. Health.*, vol. 22, no. 6, pp. 1765–1774, 2018.

[14] L. A. Sanchez-Perez et al., "Rest tremor quantification based on fuzzy inference systems and wearable sensors," *International journal of medical informatics*, vol. 114, pp. 6–17, 2018.

[15] A. Samà et al., "Determining the optimal features in freezing of gait detection through a single waist accelerometer in home environments," *Pattern Recognition Letters*, vol. 105, pp. 135–143, 2018.

[16] P. Prabhu et al., "Classification of gait signals into different neurodegenerative diseases using statistical analysis and recurrence quantification analysis," *Pattern Recognition Letters*, 2018.

[17] T. Chomiak et al., "A novel single-sensor-based method for the detection of gait-cycle breakdown and freezing of gait in Parkinson's disease," *Journal of Neural Transmission*, vol. 126, pp. 1029–1036, 2019.

[18] P. A. Perez-Toro et al., "Nonlinear dynamics and poincaré sections to model gait impairments in different stages of Parkinson's disease," *Nonlinear Dynamics*, vol. 100, no. 4, pp. 3254–3276, 2020.

[19] J. Camps et al., "Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit," *Knowledge-Based Systems*, vol. 139, pp. 119–131, 2018.

[20] F. M. Pfister et al., "High-resolution motor state detection in parkinson's disease using convolutional neural networks," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[21] J. C. Vásquez-Correa et al., "Multimodal assessment of parkinson's disease: a deep learning approach," *IEEE J. Biomed. Health.*, vol. 23, no. 4, pp. 1618–1630, 2018.

[22] J. C. Vásquez-Correa et al., "Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease," *Journal of Communication Disorders*, vol. 76, pp. 21–36, 2018.

[23] J. R. Orozco-Arroyave et al., "Apkinson: the smartphone application for telemonitoring Parkinson's patients through speech, gait, and hands movement," *Neurodegenerative Disease Management*, 2020.

[24] J. C. Vásquez-Correa et al., "Effect of acoustic conditions on algorithms to detect parkinson's disease from speech," in *ICASSP*, 2017, pp. 5065–5069.

[25] T. Arias-Vergara et al., "Unobtrusive monitoring of speech impairments of parkinson's disease patients through mobile devices," in *ICASSP*, 2018, pp. 6004–6008.

[26] J. Rusz et al., "Smartphone allows capture of speech abnormalities associated with high risk of developing parkinson's disease," *IEEE T. Neur. Sys. Reh.*, 2018.