Multimodal assessment of Parkinson's disease: a deep learning approach

J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth

***Abstract*—Parkinson's disease is a neurodegenerative disorder characterized by a variety of motor symptoms. Particularly, difficulties to start/stop movements have been observed in patients. From a technical/diagnostic point of view, these movement changes can be assessed by modeling the transitions between voiced and unvoiced segments in speech, the movement when the patient starts or stops a new stroke in handwriting, or the movement when the patient starts or stops the walking process. This study proposes a methodology to model such difficulties to start or to stop movements considering information from speech, handwriting, and gait. We used those transitions to train convolutional neural networks to classify patients and healthy subjects. The neurological state of the patients was also evaluated according to different stages of the disease (initial, intermediate, and advanced). In addition, we evaluated the robustness of the proposed approach when considering speech signals in three different languages: Spanish, German, and Czech. According to the results, the fusion of information from the three modalities is highly accurate to classify patients and healthy subjects, and it shows to be suitable to assess the neurological state of the patients in several stages of the disease. We also aimed to interpret the feature maps obtained from the deep learning architectures with respect to the presence or absence of the disease and the neurological state of the patients. As far as we know, this is one of the first works that considers multimodal information to assess Parkinson's disease following a deep learning approach.**

***Index Terms*—Parkinson's Disease, Deep Learning, Convolutional Neural Networks, Speech, Handwriting, Gait.**

I. Introduction

PARKINSON'S disease (PD) is the second most common neurodegenerative disorder in the world, and affects about 2% of people older than 65 years [1]. PD is characterized by the progressive loss of dopaminergic neurons in the midbrain producing several motor and non-motor impairments [2]. Motor symptoms include among others, bradykinesia, rigidity, resting tremor, micrographia, and different speech impairments. Non–motor symptoms include depression, sleep disorders, impaired language, and others [3]. The level and characteristics of motor impairments are currently evaluated according to the Movement Disorder Society – Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [4]. Section III of the scale contains several items to assess motor impairments. The evaluation requires the patient to be present at the clinic, which is expensive and time-consuming due to several limitations including the availability of neurologist experts in the hospital and the reduced mobility of the patients. The evaluation of motor capabilities is crucial for clinical experts to make decisions about the medication dose or therapy exercises for the patients [5]. The analysis of bio-signals such as gait, handwriting, and speech helps in objectively assessing motor symptoms of patients, providing additional and objective information to clinicians to make accurate and timely decisions about the treatment. The research community is interested in developing technology that helps the automatic evaluation of the neurological state of PD patients considering different bio-signals such as speech, handwriting, and gait.

A. Assessment of PD from speech

Speech symptoms in PD patients are grouped and typically called hypokinetic dysarthria. They include monopitch, reduced stress, imprecise consonants, and reduced loudness. One of the first observed impairments was the imprecise production of stop consonants such as /p/, /t/, /k/, /b/, /d/, and /g/ [6]. Other symptoms include reduced duration of vocalic segments and transitions, and increased voice onset time [6], [7], which may increase with the disease progression. Several studies have described speech impairments developed by PD patients in terms of different dimensions: phonation, articulation, prosody, and intelligibility [8], [9]. Phonation symptoms are related to the stability and periodicity of the vocal fold vibration. They have been analyzed in terms of perturbation measures such as jitter, shimmer, amplitude perturbation quotient, pitch perturbation quotient, and non-linear dynamics measures [10], [11]. Articulation symptoms are related to the modification of position, stress, and shape of several limbs and muscles to produce speech. These symptoms have been modeled by vowel space area, vowel articulation index, formant centralization ratio, diadochokinetic analysis (DDK), and the onset energy [8], [11], [12]. Prosody deficits are manifested as monotonocity, monoloudness, and changes in speech rate and pauses. Prosody analyses are mainly based on pitch and energy contours, and duration [13].

Besides classical feature extraction methods to model pathological speech, deep learning methods have been successfully implemented in recent years to evaluate specific phenomena in speech, including the detection and monitoring of PD [14],

J. C. Vásquez-Correa, T. Arias-Vergara, and J. R. Orozco-Arroyave are with Faculty of Engineering, University of Antioquia UdeA, Medellín, Colombia.

J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, and E. Nöth are with Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

T. Arias-Vergara is with Ludwig-Maximilians-University, Munich, Germany

B. Eskofier, is with the Machine Learning and Data Analytics Lab, Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

J. Klucken is with the Department of Molecular Neurology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

Corresponding author: jcamilo.vasquez@udea.edu.co

Manuscript received XXX, 2018; revised XXXX

[15]. These methods have improved the performance of the models compared to the results obtained with classical machine learning approaches. For instance, the "2015 Computational Paralinguistics challengE (ComParE)" [16] had one of the sub-challenges about the automatic estimation of the neurological state of PD patients according to the MDS-UPDRS-III score. The winners [14] reported a correlation of 0.65 using Gaussian processes and deep neural networks (DNN) to predict the clinical scores. In [17] it was proposed a deep learning model to assess dysarthric speech. The model aimed to predict the severity of dysarthria adding an intermediate interpretable hidden layer that contains four perceptual dimensions: nasality, vocal quality, articulatory precision, and prosody. The authors presented an interpretable output highly correlated (Spearman's correlation of up to 0.82) with subjective evaluations performed by speech and language pathologists. In [18] the authors modeled the composition of non-modal phonations in PD. The authors computed phonological posteriors using deep neural networks. Those phonological posteriors were used to predict the dysarthria level of 50 PD patients and 50 HC speakers. In [15] the authors modeled articulation impairments of PD patients with time-frequency representations (TFR) and convolutional neural networks (CNNs). The authors classified PD and HC speakers considering speech recordings in three languages: Spanish, German, and Czech, and reported accuracies from 70% to 89%, depending on the language, indicating that deep learning methods are promising to assess the speech of patients suffering from PD.

B. Assessment of PD from handwriting

PD patients show deficits in learning new movements, particularly in handwriting, patients exhibit impaired peak acceleration and stroke size, i.e., micrographia [19]. Speed in handwriting of PD patients is also reduced compared to age– and gender–matched HC subjects [20]. Impaired force amplitude and timing have also been observed [21]. In [22] the authors used a smart pen with integrated acceleration and pressure sensors to extract statistical and spectral features. The authors classified PD vs. HC subjects and reported an accuracy of 89% using an Adaboost classifier. In [23] the authors considered several machine learning methods to discriminate between PD patients and HC subjects. The authors evaluated the in–air and on–surface hand-movements with kinematics and pressure features, and reported accuracies of up to 85%.

C. Assessment of PD from gait

The most common manifestations of PD appear in gait, and typically cause disability in patients. Several works have studied the impact of PD in gait. In [24] the authors classified specific stages and motor signs of PD using the Embedded Gait Analysis using Intelligent Technology (eGaIT) system. The authors identified different stages of the disease according to the UPDRS scores. In [25] several inertial sensors attached to the lower and upper limbs were used to predict the UPDRS scores of 34 PD patients. The authors computed features related to stance time, length of the stride, and velocity of each step, and reported a Pearson's correlation coefficient of 0.60 between the estimated and real UPDRS scores. Recently, in [26] the authors proposed two novel interpretable features to assess gait impairments in PD patients: the peak forward acceleration in the loading phase and peak vertical acceleration around heel-strike. These two features encode the engagement in stride initiation and the hardness of the impact at heel-strike, respectively. The features were correlated with the UPDRS-III scores of 98 PD patients and the results indicated that the proposed features are suitable to evaluate the disease progression and loss of postural agility/stability of patients.

D. Multimodal analysis of PD

Although there are several works considering different biosignals to assess motor impairments of PD patients, most of the studies consider only one modality. Multimodal analyses, i.e., considering information from different sensors, have not been extensively studied [27]. Additionally, the robustness of the existing signal processing and classification algorithms has not been enough tested using information from the combination of multiple sensors. Although many improvements have been shown in several tasks, there is still an absence of a multimodal fusion system able to deliver an accurate prediction of the PD severity [28] and to monitor the disease progression. In [22] the authors combined information from statistical and spectral features extracted from handwriting and gait signals. The fusion of features improved the accuracy of the classification between PD and HC subjects. In previous studies [29] we also found that the combination of bio–signals improved the results regarding the assessment of the motor capabilities of the patients. The results improved in classification and regression experiments, where the capability of the model to predict the disease severity was evaluated.

E. Contribution of this study

On the basis of clinical evidence that shows difficulties of patients suffering from PD to start and stop movements [7], i.e., the transitions, and following the idea proposed in [11], this paper introduces a methodology to model such transitions in speech, handwriting, and gait signals. The aims of this work include to evaluate the neurological state of the patients, to assess specific impairments in the lower/upper limbs and muscles, and to evaluate the impact of the disease in speech. To address these aims, onset (to start voluntary movements) and offset (to stop voluntary movements) transitions are detected in speech, on-line handwriting, and gait. Speech transitions are detected when the patients start/stop the vibration of vocal folds. Transitions in handwriting are detected when the patient has the pen in the air and puts it on the tablet's surface, and gait transitions are detected when the patient starts/stops walking. These transitions are modeled considering a deep learning approach based on CNNs. Several experiments are performed to classify PD vs. HC subjects and to evaluate the neurological state of patients in several stages of the disease. Specific motor impairments in lower/upper limbs and in speech are assessed to classify the patients into three stages of the disease (initial, intermediate, and severe). We aim also to find an interpretation of the feature maps obtained from

the CNNs in each convolutional layer. We obtained state-of-art results for the classification of PD vs. HC subjects using multimodal information. As far as we know, this is one of the first studies that considers multimodal information to assess motor capabilities of PD patients using deep learning approaches. Besides the multimodal analysis, the robustness of the proposed approach is evaluated considering speech signals in three different languages: Spanish, German, and Czech. These kinds of multilingual experiments have been performed before considering classical machine learning techniques [11] but not with deep learning approaches.

II. Data

A. Multimodal data

The data contain recordings of speech, handwriting, and gait collected from 44 (29 female) PD patients and 40 HC subjects (18 female). Both groups are balanced in gender $[\chi^2(0.05) = 7.21, d = 38, p = 0.99]$. All of the subjects are Colombian Spanish native speakers. None of the participants in the HC group has history of symptoms related to PD or any other kind of movement disorder. The patients were evaluated by a neurologist expert and labeled according to the MDS-UPDRS-III scale. All the patients were recorded in ON-state. Most of them were under pharmacotherapy (unfortunately we did not have access to the data of the medication doses), which have shown to reduce the impact of speech impairments in PD patients [30]. It also improves several gait symptoms, including those assessed with the proposed approach, e.g., gait initiation and freezing of gait [31]. For handwriting, the dopaminergic medication has shown partial improvement in the kinematics of the process [20]. The three bio-signals were captured in the same session during 1 hour, distributed as follows: 15 minutes for speech, 30 minutes for gait, and 15 minutes for handwriting. Table I shows demographic information of the subjects. We divided the total MDS-UPDRS-III score into three sub-scores to analyze specific impairments in the lower limbs, upper limbs, and speech. The speech score ranges from 0 to 4 and corresponds only to one item. The sum of the scores to asses upper and lower limbs ranges from 0 to 56, corresponding to 14 items of the complete scale. The division of the items is shown in Table II. Figure 1 shows the distribution of the scores for the multimodal data.

TABLE I
General information about multimodal data. μ: average, σ: standard deviation.

	PD patients		HC subjects	
	male	female	male	female
Number of subjects	15	29	21	18
Age [years] ($\mu \pm \sigma$)	62.5 ± 9.7	57.8 ± 11.1	67.4 ± 12.8	60.5 ± 8.0
Range of age [years]	41–81	25–75	49–84	50–74
Disease duration [years] ($\mu \pm \sigma$)	8.0 ± 4.4	12.8 ± 12.4		
Range of disease duration [years]	1–15	0–43		
MDS-UPDRS-III ($\mu \pm \sigma$)	34.6 ± 22.1	36.3 ± 24.2		
Range of MDS-UPDRS-III	8–82	9–106		

Three classes are defined from each histogram to perform multi-class experiments to discriminate between initial, intermediate, and severe stages of the disease. For the complete MDS-UPDRS-III score the ranges per class are defined as follows: 0 to 25 (initial), 25 to 50 (intermediate), and higher than 50 (severe). For the sub-scores related to lower and upper limbs, the classes are defined as 0 to 10 (initial), 10 to 22 (intermediate), and higher than 22 (severe). Finally for the speech item, we consider 0 as the initial stage, 1 as the intermediate stage, and 2 or higher as the severe stage. The distribution and limits of the scores per class are shown in Figure 1. Note that one patient could be in different classes per sub-score depending on which limbs/muscles are more affected, e.g., the same patient could be in initial stage in speech, intermediate in upper limbs, and severe in lower limbs.

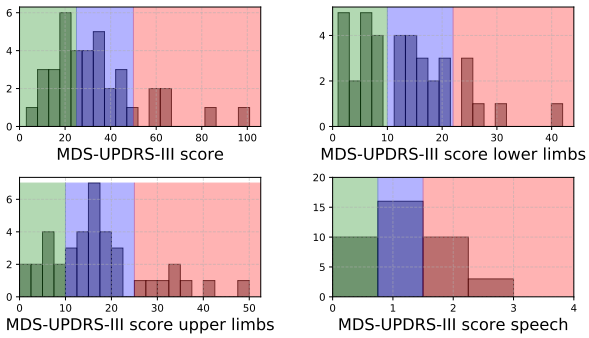

Fig. 1. Histograms for the complete MDS-UPDRS-III score and its three sub-scales for upper limbs, lower limbs, and speech. Patients in initial stage (green), patients in intermediate stage (blue), and patients in severe stage (red).

1) Recorded data: The speech of the participants was recorded with a sampling frequency of 16 kHz and 16-bit resolution. The participants pronounced six DDK exercises: the rapid repetition of the syllables /pa-ta-ka/, /pe-ta-ka/, /pa-ka-ta/, /pa/, /ta/, and /ka/. Additionally, the corpus contain read sentences, a read story of 36 words, and a monologue. Handwriting data consist of on-line drawings captured with a tablet Wacom cintiq 13-HD[1] with a sampling frequency of 180 Hz. The tablet captures six different signals: x-position, y-position, in-air movement, azimuth, altitude, and pressure. The subjects performed a total of 14 tasks divided into writing and drawing tasks (See Table III for details of the performed tasks). To give an idea of the information that can be obtained from the on-line handwriting, Figure 2 shows Archimedian spirals drawn by one HC subject and three patients in different stages of the disease (low, intermediate, and severe).

Gait signals were captured with the eGaIT system[2]. The system consists of a 3D-accelerometer (range ±6g) and a 3D gyroscope (range ±500°/s) attached to the lateral heel of the shoes [24]. Data from both foot were captured with a sampling rate of 100 Hz and 12-bit resolution. The tasks included 20 meters walking with a stop after 10 meters (2×10 walk), and 40 meters walking with a stop every 10 meters (4×10 walk).

B. Additional speech data

Besides the multimodal data, we consider three additional speech datasets with recordings in three languages: Spanish,

[1]Cintiq 13HD Graphic pen tablet for drawing http://www.wacom.com/en-us/products/pen-displays/cintiq-13-hd

[2]eGaIT - embedded Gait analysis using Intelligent Technology, http://www.egait.de/

TABLE II
DIVISION OF THE MDS-UPDRS-III SCORE INTO SUB-ITEMS FOR SPEECH, LOWER LIMBS, AND UPPER LIMBS.

Item	Description	Group	Item	Description	Group
3.1	Speech	Speech	3.10	Gait	Lower limbs
3.3b-c	Rigidity in right-left up extremities	Upper limbs	3.11	Freezing of gait	Lower limbs
3.3d-e	Rigidity in right-left low extremities	Lower limbs	3.12	Postural stability	Lower limbs
3.4	Finger tapping	Upper limbs	3.13	Posture	Lower limbs
3.5a-b	Left-Right hand movements	Upper limbs	3.14	Global spontaneity of movement	Lower limbs
3.6a-b	Pronation-Supination left-right hands	Upper limbs	3.15a-b	Postural tremor of left-right hands	Upper limbs
3.7a-b	Left-Right toe tapping	Lower limbs	3.16a-b	Kinetic tremor of left-right hands	Upper limbs
3.8a-b	Left-Right leg agility	Lower limbs	3.17a-b	Rest tremor amplitude left-right up extremities	Upper limbs
3.9	Arising from chair	Lower limbs	3.17c-d	Rest tremor amplitude left-right low extremities	Lower limbs
			3.18	Constancy of rest tremor	Upper limbs

TABLE III
HANDWRITING TASKS PERFORMED BY THE PARTICIPANTS.

Writing tasks	Drawing tasks
The name	A circle
The ID number	A cube
The numbers from 0 to 9	Two rectangles
A template sentence*	A house
A free sentence	A diamond
The signature	The Rey-Osterrieth figure
	A spiral following a template
	A free spiral

*Template sentence: *El abecedario es a b c d e f g h i j k l m n o p q r s t u v w x y z*, which translates: *The alphabet is a b c d e f g h i j k l m n o p q r s t u v w x y z*

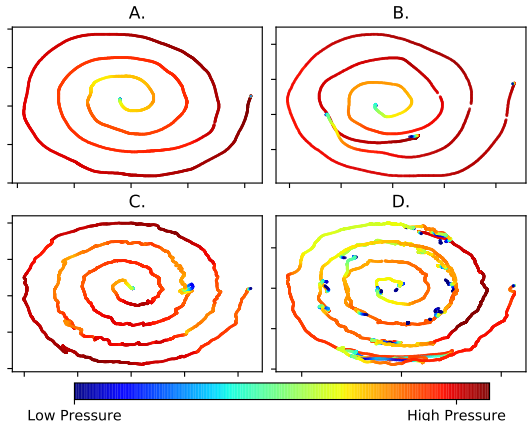

Fig. 2. Spiral drawn by: A) HC subject (male, 41 years old); B) PD patient in low state (male, 59 years old, MDS-UPDRS=8); C) PD patient in intermediate state (female, 59 years old, MDS-UPDRS=33); and D) PD patient in advance state (female, 73 years old, MDS-UPDRS=64).

German, and Czech with the aim to evaluate the robustness of deep neural networks when considering speech signals of PD patients and HC subjects in different languages. Table IV summarizes the information of each database.

1) Spanish: The corpus considered here is the PC-GITA database [32]. The data contain speech recordings of 50 PD (25 women) and 50 HC (25 women) Colombian Spanish native speakers. All of them are balanced in age $[t(0.05) = -0.2878, p = 0.99]$. Twenty of these patients participated also in the collection of the multimodal data. All of the speakers pronounced the same speech tasks that were considered in the multimodal data. All of the patients were recorded in ON state, i.e., no more than three hours after their morning medication, and were evaluated by the same neurologist that participated in the collection of the multimodal data.

TABLE IV
GENERAL INFORMATION ABOUT THE SPEECH DATA IN EACH LANGUAGE. **PD**: PARKINSON'S DISEASE. **HC**: HEALTHY CONTROLS. μ: AVERAGE, σ: STANDARD DEVIATION.

	PD patients		HC subjects	
	male	female	male	female
Spanish				
Number of subjects	25	25	25	25
Age [years] ($\mu \pm \sigma$)	61.3 ± 11.4	60.7 ± 7.3	60.5 ± 11.6	61.4 ± 7.0
Range of age [years]	33–81	49–75	31–86	49–76
Disease duration [years] ($\mu \pm \sigma$)	8.7 ± 5.8	12.6 ± 11.6		
Range of disease duration [years]	1–20	1–43		
MDS-UPDRS-III ($\mu \pm \sigma$)	37.8 ± 22.1	37.6 ± 14.1		
Range of MDS-UPDRS-III	6–93	19–71		
German				
Number of subjects	47	41	44	44
Age [years] ($\mu \pm \sigma$)	66.7 ± 9.0	66.1 ± 9.0	63.8 ± 14.0	62.6 ± 13.9
Range of age [years]	44–82	42–84	26–83	28–85
Disease duration [years] ($\mu \pm \sigma$)	6.5 ± 5.8	6.8 ± 5.9		
Range of disease duration [years]	1–19	1–30		
UPDRS ($\mu \pm \sigma$)	22.1 ± 10.9	23.3 ± 10.8		
Range of UPDRS	5–43	6–55		
Czech				
Number of subjects	20	0	16	0
Age [years] ($\mu \pm \sigma$)	61.0 ± 12.0	–	61.8 ± 13.3	–
Range of age [years]	34–83	–	36–80	–
Disease duration [years] ($\mu \pm \sigma$)	2.4 ± 1.7	–		
Range of disease duration [years]	0–7	–		
UPDRS ($\mu \pm \sigma$)	17.9 ± 7.3	–		
Range of UPDRS	5–32	–		

2) German: The German data contain recordings of 88 PD patients (41 women) and 88 HC subjects (44 women). The speakers are balanced in age $[t(0.05) = -2,056, p = 0,02]$. The speakers performed several speech tasks, including the repetition of /pa-ta-ka/. Further details of this corpus can be found in [13].

3) Czech: The Czech data are formed with recordings of 20 PD patients and 15 HC subjects. All of them are men. The patients were newly diagnosed with PD, and none of them had been medicated before or during the recording session. The speakers are balanced in age $[t(0.05) = 0.31, p = 0.31]$. The speakers performed several speech tasks, including the repetition of /pa-ta-ka/. Further details about this corpus can be found in [33].

III. DETECTION OF THE START/STOP MOVEMENT

The transition movements in speech, handwriting, and gait are detected individually upon each bio-signal to model difficulties of the patients to start/stop the movement.

A. Transitions in speech

A transition in speech occurs when the speaker starts or stops the vocal fold vibration. We detected the transition

from unvoiced to voiced segments (onset) and from voiced to unvoiced (offset). Those transitions are produced by the combination of different sounds during the production of continuous speech. Offsets and onsets are segmented according to the presence of the fundamental frequency F_0 using Praat. Once the borders are detected, 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 TFR using the short-time Fourier transform (STFT). The TFR is used as input to the deep learning architecture. Figure 3 shows the difference in the onsets between one HC subject and three patients in different stages of the disease (low, intermediate, and severe). Note that the HC speaker clearly defines the transition, conversely the patients are not able to produce clean transitions.

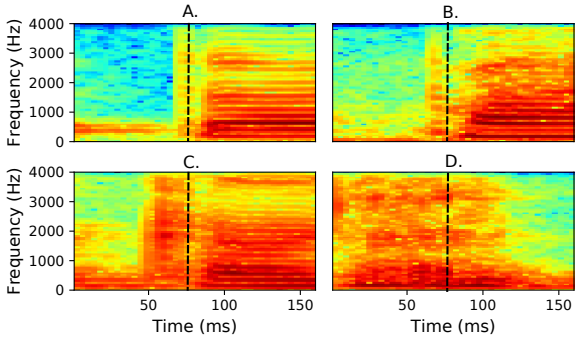

Fig. 3. STFT of an onset produced by: A) a 75 years old female HC subject; B) a 72 years old female PD patient in low state of the disease (MDS-UPDRS=19); C) a 73 years old female PD patient in intermediate state (MDS-UPDRS=38); and D) a 75 years old female PD patient in severe state (MDS-UPDRS=52). All figures correspond to the syllable /ka/.

B. Transitions in gait

Gait transitions appear when the patient starts (onset) or stops (offset) walking. These transitions are segmented according to the presence of the fundamental frequency of the signal, which is related to the acceleration of each stride. In addition, an energy-based threshold is considered to improve the robustness in the detection of onsets and offsets. Similar to speech, once a border is detected frames of 3 s are considered to each side of the border guaranteeing at least 3 quasi-periods in each "chunk" of signal. The STFT is computed upon the onsets and offsets and it is used as input for the deep learning model. Figure 4 shows the difference in the onset produced by one HC subject and three patients in different stages of the disease (low, intermediate, and severe). Theses images are extracted from the z-axis gyroscope signal from the left foot. The six signals captured with the inertial sensors are used as inputs to the deep learning architecture.

C. Transitions in handwriting

Transitions in handwriting occur when a starting point of a stroke is detected (onset), or when the pen takes-off the surface of the tablet after drawing a stroke (offset). Once each border is detected, segments of 200 ms are taken to the left and to

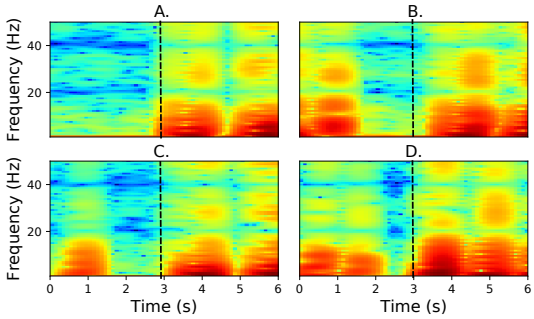

Fig. 4. STFT of a gait onset produced by: A) a 68 years old male HC subject; B) a 62 years old female PD patient in low state (MDS-UPDRS=19); C) a 65 years old male PD patient in intermediate state (MDS-UPDRS=43); and D) a 57 years old male PD patient in severe state (MDS-UPDRS=58). All figures correspond to the 2×10 task.

the right of the six signals captured with the tablet: horizontal movement (x), vertical movement (y), distance between the surface and the pen (z), azimuth angle, altitude angle, and pressure of the pen. Figure 5 shows the handwriting onset of one HC subject and three patients in different stages of the disease (low, intermediate, and severe). Note that the dynamics of the z-axis (black lines) is different for PD patients and HC subjects before starting the stroke (the first 0.5 s of the figure). Note that the resting tremor in the PD patients is clearly observed, especially for the PD patient in Figure 5C, where oscillations around 7 Hz are observed when the pen is in the air. Complementary material with figures for all PD and HC subjects can be found on-line[3].

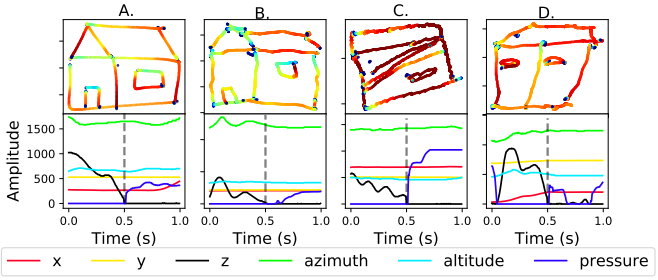

Fig. 5. Handwriting onset produced by: A) a 68 years old male HC subject; B) a 48 years old male PD patient in low state (MDS-UPDRS=13); C) a 41 years old male PD patient in intermediate state(MDS-UPDRS=27); and D) a 75 years old female PD patient in severe state (MDS-UPDRS=108).

IV. Deep Learning Architectures

Architectures based on CNNs are considered as the deep learning models in this study for several reasons: (1) the data modalities considered here are in the form of multiple arrays e.g., 2D speech and gait spectrograms, and 1D handwriting signals, which makes CNNs the most suitable deep learning architectures to process such information; (2) we aim to take advantage of four key aspects of CNNs to process the biosignals considered in this study: local connections, shared weights, pooling, and the use of many layers; (3) CNNs are

[3]https://github.com/jcvasquezc/images_deep_transition

able to detect different local motifs that may appear in the multiple dimension array due to high correlations between neighbor values [34]. This concept would allow to detect for instance spectral bands with more energy density in speech or gait to discriminate between PD patients and HC subjects.

A. Convolutional neural networks

A CNN is a variant of the standard neural networks. Instead of using fully connected hidden layers, the CNN introduces a structure that consists of alternating convolution and pooling layers. CNNs have been used in several tasks of speech and audio processing like classification of pathological speech [15], detection of events in audio, speech recognition, and others. CNNs are designed to process data from multiple arrays, for instance a color image formed by three channels (RGB), or two-dimensional arrays that correspond to TFR of audio signals. CNNs introduce a structure formed by alternating convolutional filters and pooling layers instead of the fully connected layers of a DNN. The input of a CNN is a tensor $\mathbf{X} \in \mathbb{R}^{p\times q\times c}$, where p, q and c can be the number of vertical pixels, horizontal pixels, and channels of an RGB image, respectively. The convolution is performed between the input $\mathbf{X}$ and a weight tensor $\mathbf{W} \in \mathbb{R}^{n\times n\times d}$ producing a hidden representation $\mathbf{H} \in \mathbb{R}^{(p-n+1)\times(q-n+1)\times d}$ that contains the extracted features from the input. n is the order of the convolutional filter and d is the number of feature maps in the convolutional layer. After the convolution, a pooling layer is applied to remove variability that may appear due to external factors like the speaking style or channel distortion. The last layer of a CNN corresponds to a fully connected layer with h hidden units followed by a sigmoid activation function to make the final decision of whether the TFR corresponds to a PD patient or a HC speaker. In this study, several CNNs are used to extract information from speech, handwriting and gait. For the speech and gait signals, two-dimensional (2D) CNNs are trained to process the TFR created with the STFT of the transitions, as in previous studies [15]. As the speech recordings are monophonic, in this case only one channel is considered in the input of the CNNs. For gait analysis the input consists of $c = 12$ channels that contain signals of the accelerometer and gyroscope in the x, y, and z-axes of the left and right foot. For on-line handwriting, we consider a 1D CNN with $c = 16$ channels that include information of the transition from in-air to on-surface movement, or vice-versa. In this case the inputs to the CNN consist of the raw data of eight signals: x–position, y–position, z–position, pressure of the pen, azimuth angle, altitude angle, on-surface trajectory (r), and angle of the trajectory (θ). All of them are captured in the transitions. The derivatives of these data are also included to complete the 16 channels. Table V summarizes the inputs received by the CNN for each bio-signal. A STFT with 128 points is computed for speech and gait, forming the 65 frequency indexes in the input. Frames of 16 ms with a time-shift of 4 ms are considered for the STFT in the speech signals, forming a total of 40 frames. The frame size in gait is 200 ms with a time-shift of 100 ms, forming 60 frames. Note that the number of inputs for gait is much larger than the inputs for speech and handwriting, which gives an idea about the complexity of the CNNs for each bio-signal.

TABLE V
NUMBER OF INPUTS OF THE CNNS FOR SPEECH, GAIT, AND HANDWRITING SIGNALS. c: NUMBER OF CHANNELS.

Input signal	Convolution	Input size	c	Num. inputs
Speech	2D	40 ×65	1	2600
Gait	2D	60 ×65	12	46800
Handwriting	1D	180 ×1	16	2880

Figure 6 shows the CNN architecture used in this study. Figure 6A depicts a 2D-CNN with two convolutional and max–pooling layers followed by a fully connected layer that receives the TFRs as input from speech or gait. Figure 6B illustrates a 1D-CNN with 2 convolutional and pooling layers to process the raw information of the transitions in handwriting.

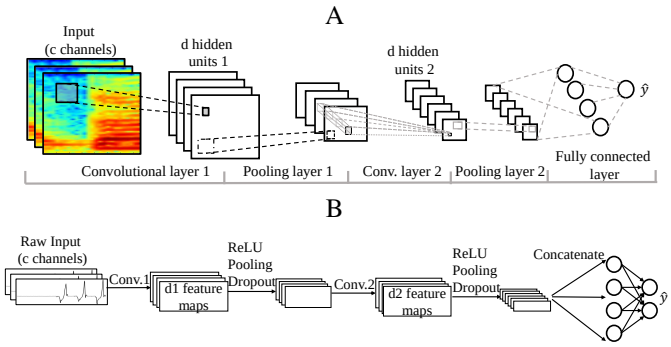

Fig. 6. CNN architectures implemented in this study.

The CNNs are trained using the stochastic gradient descent (SGD) algorithm. The cross–entropy between training labels y and the model predictions $\widehat{y}$ is used as the loss function for classification. This cost function is related to the negative log-likelihood of the model. The root mean square propagation is considered as a mechanism to adapt the learning rate in each iteration t for each parameter of the network. The method divides the learning rate η by an exponentially decaying average of squared gradients using Equations 1 and 2 [35], where g' indicates the derivative of the parameters Θ in the t-th iteration.

$$G(\Theta)^{(t)} = 0.9G(\Theta)^{(t-1)} + 0.1g'(\Theta^{(t)})^2 \tag{1}$$

$$\eta^{(t)} = \frac{\eta^{(t-1)}}{\sqrt{G(\Theta)^{(t)}}} \tag{2}$$

Additionally, rectifier linear (ReLU) activation functions are used in the convolutional layers, and dropout is included in the training stage to avoid over–fitting. The architecture of the CNN implemented in this study consists of four convolutional layers, two max-pooling layers, dropout to regularize the weights, and two fully connected hidden layers followed by the output layer to make the final decision using a sigmoid activation function. Details of this architecture are summarized in Table VI.

TABLE VI
CNN ARCHITECTURE FOR MULTIMODAL ANALYSIS OF PD.

Input (Speech; handwriting; gait)
Convolutional layer 1 (ReLU unit)
Convolutional layer 2 (ReLU unit)
Max-pooling layer 1
Dropout 1
Convolutional layer 3 (ReLU unit)
Convolutional layer 4 (ReLU unit)
Max-pooling layer 2
Dropout 2
Fully connected hidden layer (ReLU unit)
Fully connected hidden layer (ReLU unit)
Output layer (Sigmoid unit)

B. Fusion

Individual CNNs are trained for each modality, afterwards multimodal assessment is performed by combining the three bio-signals in 3 steps: (1) the feature maps from the last hidden layer of each CNN are averaged across the different tasks and transitions of a given subject. The aim is to form one feature vector with information of all tasks per subject and per bio-signal; (2) the embeddings obtained from the three bio-signals are concatenated to form a multimodal vector per subject; and (3) the created feature vectors are used to classify PD patients and HC subjects using a radial basis SVM.

C. Baseline

Conventional feature sets and traditional machine learning methods from related studies are considered to compute the baseline. The speech signals are modeled with the 88 features of the extended Geneva minimalistic acoustic parameter set (EGeMAPS) [36], which are extracted using the OpenSMILE toolkit [37]. Handwriting strokes are modeled with kinematics features based on the trajectory, velocity, and pressure of the pen, which were used in previous studies [23], [29]. Gait features include kinematics measures based on the length of the stride, velocity of each step, swing time, and stance time [24], [29]. All features are classified using a radial basis SVM. The fusion baseline is based on the early fusion approach with features of the three bio-signals.

D. Validation

The experiments are validated with the following strategy: 80% of the data are used for training, 10% are used to optimize the hyper-parameters, i.e., development set, and the remaining 10% of the data are used for test. The process is repeated 10 times with different partitions of the test set to guarantee that every participant is only tested once.

The hyper-parameter tuning is performed with a Bayesian optimization approach [38] due to the large number of hyper-parameters that needs to be optimized. Bayesian optimization is one of the sequential model-based optimization (SMBO) algorithms. The hyper-parameter tuning is an optimization problem, where we find the hyper-parameters that maximize the performance of the model on the development set. SMBO algorithms use previous observations of a loss function f, to determine the next (optimal) point to sample f. The Bayesian optimization assumes that the loss function f can be described by a Gaussian Process (GP). The GP induces a posterior distribution over the loss function f that is analytically tractable, which allows us to update f, after we have computed the loss for a new set of hyper-parameters. The Expected Improvement (EI) is used as the optimization function for the Bayesian optimization algorithm. The EI is the expected probability that a new set of hyper-parameters will improve the current best observation. EI is defined as $\mathrm{EI}(\beta) = \mathbb{E}[\max\{0, f(\beta) - f(\widehat{\beta})\}]$, where β is the current set of hyper-parameters and $\widehat{\beta}$ is the current optimal set of hyper-parameters. EI will give us the point that in expectation improves the most upon f. The Bayesian optimization algorithm can be summarized according to the following steps:

1) Given observed values of $f(\beta)$, update the posterior expectation of f using the GP model.
2) Find β_{new} that maximizes $\mathrm{EI}(\beta)$.
3) Compute the loss function for $f(\beta_{\mathrm{new}})$.

We use the accuracy in the development set as the optimization function $f(\beta)$, and the hyper-parameters set β is formed with the filter size of each convolutional layer of the CNN (n_i), the number of feature maps in each convolutional layer (d_i), the number of hidden units in the fully connected layers h_1 and h_2, the initial learning rate η, and the probability of dropout. The range of the hyper-parameters to be optimized is shown in Table VII. In addition a batch-size of 64 samples and a total of 150 epochs are considered.

TABLE VII
RANGE OF THE HYPER-PARAMETERS USED TO TRAIN THE CNNS.

Hyper-parameter	Values
Filter size convolutional layers	$\{3, 5, 7\}$
Depth of convolutional layers	$\{4, 8, 16, 32, 64\}$
Hidden units in fully connected layers	$\{16, 32, 64, 128\}$
Learning rate	$\{0.0001, 0.0005, 0.001\}$
Probability of dropout	$\{0.1, 0.2 \cdots 0.9\}$

V. EXPERIMENTS AND RESULTS

A. Classification of PD patients vs. HC subjects considering multimodal data

The results considering speech, handwriting, and gait are shown in Table VIII, which includes accuracy in the development and test sets, area under the receiving operating characteristic curve (AUC) and number of parameters in the CNN. The best results are obtained with the fusion of the three bio-signals (accuracy of 97.6%). This result exceeds those obtained with each bio-signal separately and with early-fusion (the baseline). Results obtained with traditional features extracted per bio-signal are also included in Table VIII. Note that the results obtained with the proposed approach in speech and gait exceed those obtained in the corresponding baselines in 17.8% and 17.3%, respectively.

Table VIII shows the reduction of the accuracies obtained in development and test. In speech the decrease ranges between 6.7 and 15.6%. The results in gait are relatively more stable

TABLE VIII
MULTIMODAL CLASSIFICATION OF PD PATIENTS AND HC SUBJECTS. **ACC. TEST**: ACCURACY IN THE TEST SET, **ACC. DEV.**: ACCURACY IN THE DEVELOPMENT SET, **AUC**: AREA UNDER THE ROC CURVE, **N.**: NUMBER OF PARAMETERS IN THE CNN.

Bio-signal	Acc. Test	Acc. Dev.	AUC	N.
Speech baseline	74.5±1.7	77.0±2.4	0.841	
Speech onset	92.3±12.3	99.4±0.7	0.963	140055
Speech offset	83.5±6.6	99.1±0.7	0.925	135389
Gait baseline	63.0±8.9	66.0±3.1	0.725	
Gait onset	80.3±10.3	83.3±8.9	0.878	326977
Gait offset	78.8±16.0	87.8±5.1	0.901	1231016
Handwriting baseline	67.1±4.2	67.7±1.7	0.725	
Handwriting onset	60.4±3.5	95.7±4.0	0.634	142517
Handwriting offset	66.5±5.5	98.1±1.7	0.699	255560
Fusion baseline	89.0±7.8	87.8±3.1	0.944	
Fusion onset	**97.6±2.9**	98.8±0.6	0.988	609549
Fusion offset	84.3±5.8	86.0±1.4	0.890	1621965

with a decrease ranging from 0.8 to 9.0%. Handwriting seems to be the least robust for generalization purposes. The difference in the accuracy obtained in development and test ranges between 9.5 and 35.3%. It is interesting to note that the accuracies in development obtained with gait are lower than those with speech and handwriting. This fact can be explained due to the difference in the number of transitions, which limits the amount of information considered to generate the proposed model. In speech and handwriting, there are several (more than 5) transitions, while in gait there is only one transition in the case of the 2×10 task, and three in the case of the 4×10 task. Further experiments, considering tasks with more transitions, e.g., heel-toe taping, are required to validate this hypothesis. The only relatively high difference between the results for onset and offset is observed in speech. Such a difference could be likely explained because the DDK tasks, e.g., rapid repetition of the syllables /pa-ta-ka/, are mainly designed to assess the capability of speakers to perform onsets [39]. This behavior was also observed in previous experiments [15]. Finally, Table VIII includes the number of required parameters in the CNNs per modality. Note that gait is the modality that requires the largest number. This is expected because gait signals have the largest number of inputs, as it was shown in Table V. In order to show the results in a more compact way, Figure 7 shows the ROC curves for the best results of each modality. It can be observed that the performance in speech and gait exceeds the results obtained with handwriting.

B. Classification of PD patients vs. HC subjects considering speech signals in different languages

The generalization capability of the proposed approach is tested in several cross-language experiments. In this case only the DDK exercises of the Spanish, German, and Czech datasets are considered. The speech recordings of the three languages were re-sampled to 16kHz. CNNs were trained with features extracted from onsets/offsets of recordings of one language and tested upon recordings of the other two languages separately. Additionally, the improvement of the accuracy is analyzed when moving portions of the data in the target language to the data in the training set. The recordings

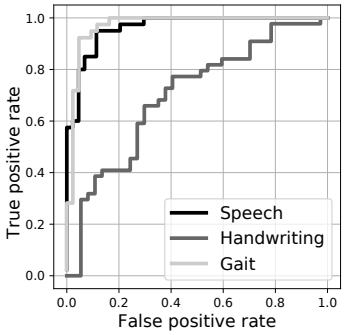

Fig. 7. ROC curves for the classification of PD patients vs. HC subjects using speech, handwriting, and gait.

of the target language are included in the test set and excluded from the training set to avoid bias. The process was repeated incrementally from 0% to 90% in steps of 10%. The results are depicted in Figure 8. Each point corresponds to the result of the aforementioned process.

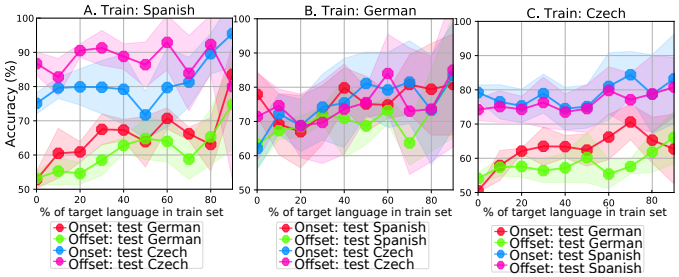

Fig. 8. Classification of PD and HC subjects in the cross-language experiments. A) Train with Spanish and test on German and Czech; B) Train with German and test on Spanish and Czech; and C) Train with Czech and test on German and Spanish.

The results obtained with the onsets are in most of the cases slightly higher than those obtained with the offsets. This behavior supports what we have observed in the experiments with multimodal data. Although the proposed approach is based on DDK exercises, which in theory are language independent, note the influence of language when no data from the target language is added to the training set, especially when the test is in the German data (Figures 8A and 8C). The language influence is reduced when moving portions of the data in the target language to the data in the training set, especially when the system is trained with Spanish utterances and tested with German recordings (Figure 8A), and when the train set is Czech and the test set is German (Figure 8C). In general, the results indicate that the proposed approach is robust against different languages, and that the DDK tasks seem to be appropriate to assess motor speech deficits in different languages. Further experiments with sentences, read texts, and spontaneous speech signals are required to address other research questions like the influence of the language in the disease manifestation and progression [40].

C. Analysis of hidden layers of CNNs

Figure 9 shows the output of the second and fourth convolutional layers of the CNN trained with the onsets of the DDK

tasks. Four feature maps are computed for the second layer, and eight for the fourth. Note that the border in the transition is more evident in the hidden layers than in the input, which may be explained due to the max pooling layer that removes non-relevant information from the spectrograms.

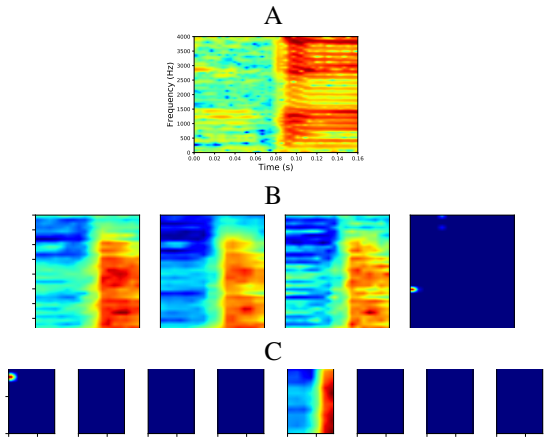

Fig. 9. Output of the CNN for the convolutional layers trained with speech onsets extracted from the DDK tasks. A) Input of the CNN; B) Outputs of the second layer; and C) Outputs of the fourth layer.

The output of the hidden layer for the gait signals is shown in Figure 10. The 12 input channels depicted in Figure 10A are transformed into the eight feature maps shown in Figure 10C in the fourth convolutional layer, forming embeddings that contain the most suitable information to classify PD patients and HC subjects. Some of the outputs of the hidden layers of the CNN are "turned-off", due to the regularization effect of the dropout, indicating that not all of the feature maps are necessary to make the final decision.

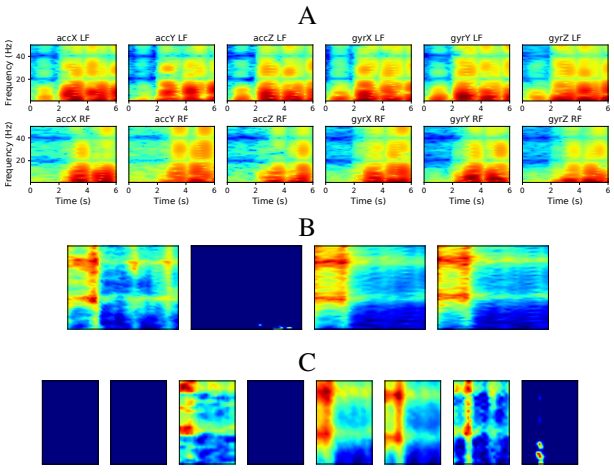

Fig. 10. Output of the CNN for each convolutional layer trained with onsets from gait. A) 12 input channels of the CNN; B) Outputs of the second layer; and C) Outputs of the fourth layer. **acc**: signals from the accelerometer, **gyr**: signals from the gysrocope, **RF**: right foot, **LF**: left foot.

The statistical difference of the computed feature maps between HC subjects and PD patients for speech, gait, and handwriting is evaluated with Kruskal-Wallis H-tests. The aim is to find which are the most discriminating feature maps and hidden layers to classify PD patients vs. HC subjects. This knowledge can help in finding possible interpretations about local features learned by the CNN in each layer, and in understanding how those features are related to the presence of the disease. The Kruskal-Wallis H-test is a non-parametric method for testing whether samples are originated from the same distribution. In this experiment it is used to evaluate the null hypothesis that the medians of the population of the tested groups are equal. The results are shown in Table IX. The second convolutional layer from speech rejects the null hypothesis for almost all of the feature maps (p-val$<$0.05). Some of the outputs of the fourth layer for speech signals show also significant differences between the PD and HC subjects. For gait signals, non of the feature maps of the hidden layers provide significant difference between the PD and HC subjects. This fact could be explained by two reasons: (1) the number of transitions in gait is much smaller than those observed in handwriting or speech, and (2) the number of inputs in the CNN for gait is much higher than those in speech and handwriting, as it was observed in Table V. These results also indicate that the fully connected hidden layers after the convolutional layers are those that provide the information to discriminate between PD and HC subjects. For handwriting, some of the features from the fourth layer have significant difference between PD and HC subjects, which supports the convenience of using those features learned by the CNN in that layer for the classification problem.

TABLE IX
KRUSKAL-WALLIS TEST TO EVALUATE THE STATISTICAL DIFFERENCE BETWEEN HC SUBJECTS AND PD PATIENTS IN THE FEATURES LEARNED BY THE CNNS FOR SPEECH, GAIT, AND HANDWRITING IN THE CONVOLUTIONAL LAYERS 2 AND 4.

	Convolutional layer 2			**Convolutional layer 4**		
	Speech	**Gait**	**Handwriting**	**Speech**	**Gait**	**Handwriting**
	p-val	p-val	p-val	p-val	p-val	p-val
feature map 1	**0.03**	0.14	0.23	$<$**0.05**	0.71	0.39
feature map 2	$<$**0.05**	0.96	0.79	$<$**0.05**	0.52	0.33
feature map 3	**0.01**	0.32	0.71	0.31	0.39	0.61
feature map 4	0.06	0.86	0.99	0.66	0.29	0.47
feature map 5			0.77	0.93	0.48	**0.01**
feature map 6			**0.02**	$<$**0.05**	0.75	0.62
feature map 7			0.49	0.08	0.84	0.98
feature map 8			0.82	$<$**0.05**	0.73	**0.04**

D. Assessment of the neurological state

Two experiments are performed to assess the neurological state of the patients. The first one aims to classify the patients into different stages of the disease according to the complete MDS-UPDRS-III scale. A total of four classes are defined according to the score assigned by the neurologist. This experiment is performed with each modality separately and with their combination. The results are shown in terms of the confusion matrices in Table X for speech, Table XI for handwriting, Table XII for gait, and Table XIII for the combination. The unweighted average recall (UAR) is computed as the global performance score. It is used to avoid bias due to the unbalance in the groups and it can be interpreted as an average ratio of the true positives per class.

The second experiment aims to classify the patients considering only those items of the MDS-UPDRS-III scale that are intended to evaluate those specific limbs and muscles of the

body that are involved in gait, handwriting, and speech. Only those items of the total MDS-UPDRS-III scale that evaluate the motor capability of the lower limbs, upper limbs, and speech are considered to assess gait, handwriting and speech deficits, respectively. The distribution of these three groups in the total scale is shown in Table II. The patients are grouped into three classes according to the specific sub-scales and they are classified separately. The division of the sub-scales is summarized in Figure 1. The results of the experiment are also shown in the confusion matrices of Table X for speech, Table XI for handwriting, and Table XII for gait.

TABLE X
CLASSIFICATION OF HC SUBJECTS AND PD PATIENTS IN THREE STAGES OF THE DISEASE USING SPEECH SIGNALS. $\mathbf{PD}_x$: PATIENTS IN LOW, INTERMEDIATE, AND SEVERE STATE ACCORDING TO THE MDS-UPDRS-III SCORE. RESULTS IN % AND ABSOLUTE VALUES (IN PARENTHESIS)

MDS-UPDRS-III score								
	Speech onset UAR=37.8%				Speech offset UAR=37.2%			
	HC	PD_1	PD_2	PD_3	HC	PD_1	PD_2	PD_3
HC	**84.6 (33)**	5.1 (2)	10.3 (4)	0.0 (0)	**82.0 (32)**	5.1 (2)	10.3 (4)	2.6 (1)
PD_1	64.3 (9)	**0.0 (0)**	35.7 (5)	0.0 (0)	78.6 (11)	**0.0 (0)**	21.4 (3)	0.0 (0)
PD_2	55.6 (10)	11.1 (2)	**33.3 (6)**	0.0 (0)	66.7 (12)	0.0 (0)	**33.3 (6)**	0.0 (0)
PD_3	16.7 (1)	16.7 (1)	33.3 (2)	**33.3** (2)	16.7 (1)	0.0 (0)	50.0 (3)	**33.3 (2)**
MDS-UPDRS-III sub-score (speech item)								
	Speech onset UAR=54.9%				Speech offset UAR=45.4%			
	HC	PD_1	PD_2	PD_3	HC	PD_1	PD_2	PD_3
HC	**92.3 (36)**	0.0 (0)	5.1 (2)	2.6 (1)	**94.9 (37)**	0.0 (0)	0.0 (0)	5.1 (2)
PD_1	10.0 (1)	**50.0 (5)**	20.0 (2)	20.0 (2)	10.0 (1)	**20.0 (2)**	50.0 (5)	20.0 (2)
PD_2	43.8 (7)	6.2 (1)	**31.2 (5)**	18.8 (3)	12.5 (2)	6.3 (1)	**43.7 (7)**	37.5 (6)
PD_3	0.0 (0)	23.0 (3)	30.8 (4)	**46.2 (6)**	15.4 (2)	15.4 (2)	46.2 (6)	**23.0 (3)**

TABLE XI
CLASSIFICATION OF HC SUBJECTS AND PD PATIENTS IN THREE STAGES OF THE DISEASE USING HANDWRITING SIGNALS. $\mathbf{PD}_x$: PATIENTS IN LOW, INTERMEDIATE, AND SEVERE STATE ACCORDING TO THE MDS-UPDRS-III SCORE. RESULTS IN % AND ABSOLUTE VALUES (IN PARENTHESIS)

MDS-UPDRS-III score								
	Handwriting onset UAR=54.9%				**Handwriting offset UAR=51.9%**			
	HC	PD_1	PD_2	PD_3	HC	PD_1	PD_2	PD_3
HC	**94.9 (37)**	2.6 (1)	0.0 (0)	2.6 (1)	**97.4 (38)**	2.6 (1)	0.0 (0)	0.0 (0)
PD_1	35.7 (5)	**35.7 (5)**	21.4 (3)	7.1 (1)	35.7 (5)	**21.4 (3)**	28.6 (4)	14.3 (2)
PD_2	22.2 (4)	16.7 (3)	**55.6 (10)**	5.6 (1)	11.1 (2)	16.7 (3)	**55.6 (10)**	16.7 (3)
PD_3	33.3 (2)	16.7 (1)	16.7 (1)	**33.3 (2)**	0.0 (0)	16.7 (1)	50.0 (3)	**33.3 (2)**
MDS-UPDRS-III sub-score (upper limbs)								
	Handwriting onset UAR=50.9%				**Handwriting offset UAR=49.7%**			
	HC	PD_1	PD_2	PD_3	HC	PD_1	PD_2	PD_3
HC	**94.9 (37)**	2.6 (1)	2.6 (1)	0.0 (0)	**92.3 (36)**	5.1 (2)	2.6 (1)	0.0 (0)
PD_1	30.0 (3)	**20.0 (2)**	50.0 (5)	0.0 (0)	40.0 (4)	**10.0 (1)**	40.0 (4)	10.0 (1)
PD_2	14.3 (3)	4.8 (1)	**76.2 (16)**	4.8 (1)	19.0 (4)	4.8 (1)	**71.4 (15)**	4.8 (1)
PD_3	0.0 (0)	0.0 (0)	87.5 (7)	**12.5 (1)**	0.0 (0)	0.0 (0)	75.0 (6)	**25.0 (2)**

The classification according to the total MDS-UPDRS-III score indicates that the highest UAR values are obtained with handwriting onsets and offsets, which can be explained because the high number of transitions that appear during the writing process, thus it can be expected to find more information in this modality than in the other two. The lowest UAR values are obtained with the speech signals, which was expected considering that the MDS-UPDRS-III scale only considers one item related to speech (see Table II), thus to classify groups according to the complete MDS-UPDRS-III scale considering only speech signals is a very difficult (and to some extent unfair) problem. The confusion matrices indicate that HC subjects are accurately classified compared

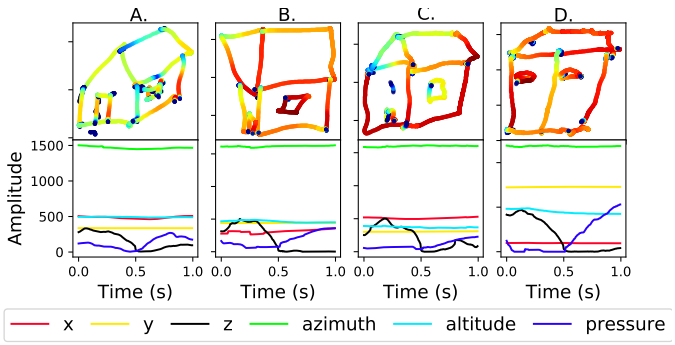

Fig. 11. Examples of drawings and onset transitions detected for miss-classified subjects: A) PD patient in low state (PD_1) detected as HC subject; B) HC classified as PD patient in low state (PD_1); C) Patient in severe state (PD_3) classified as HC; and D) Patient in severe state (PD_3) classified as PD patient in intermediate state (PD_2).

TABLE XII
CLASSIFICATION OF HC SUBJECTS AND PD PATIENTS IN THREE STAGES OF THE DISEASE USING GAIT SIGNALS. $\mathbf{PD}_x$: PATIENTS IN LOW, INTERMEDIATE, AND SEVERE STATE ACCORDING TO THE MDS-UPDRS-III SCORE. RESULTS IN % AND ABSOLUTE VALUES (IN PARENTHESIS)

MDS-UPDRS-III score								
	Gait onset UAR=48.6%				**Gait offset UAR=50.9%**			
	HC	PD_1	PD_2	PD_3	HC	PD_1	PD_2	PD_3
HC	**97.4 (38)**	0.0 (0)	2.6 (1)	0.0 (0)	**94.9 (37)**	0.0 (0)	5.1 (2)	0.0 (0)
PD_1	35.7 (5)	**35.7 (5)**	28.6 (4)	0.0 (0)	28.6 (4)	**14.3 (2)**	57.1 (8)	0.0 (0)
PD_2	33.3 (6)	16.7 (3)	**44.4 (8)**	5.6 (1)	22.2 (4)	11.1 (2)	**61.1 (11)**	5.6 (1)
PD_3	33.3 (2)	16.7 (1)	33.3 (2)	**16.7 (1)**	16.7 (1)	0.0 (0)	50.0 (3)	**33.3 (2)**
MDS-UPDRS-III sub-score (lower limbs)								
	Gait onset UAR=46.8%				**Gait offset UAR=37.6%**			
	HC	PD_1	PD_2	PD_3	HC	PD_1	PD_2	PD_3
HC	**97.4 (38)**	2.6 (1)	0.0 (0)	0.0 (0)	**89.7 (35)**	2.6 (1)	2.6 (1)	5.1 (2)
PD_1	75.0 (12)	**6.3 (1)**	12.5 (2)	6.3 (1)	50.0 (8)	**12.5 (2)**	31.3 (5)	6.3 (1)
PD_2	18.8 (3)	12.5 (2)	**50.0 (8)**	18.8 (3)	50.0 (8)	12.5 (2)	**31.3 (5)**	6.3 (1)
PD_3	16.7 (1)	0.0 (0)	50.0 (3)	**33.3 (2)**	33.3 (2)	16.7 (1)	33.3 (2)	**16.7 (1)**

to patients in different stages, i.e., the proposed approach has a high specificity. In addition, patients in the first stage of the disease (PD_1) are miss-classified mainly as HC (some of the cases are mis-classified as PD_2), which is consistent with the disease progression. Patients in a severe stage are more commonly miss-classified as patients in the intermediate than patients in the first stage or HC subjects. Figure 11 shows some examples of drawings and transitions of miss-classified handwritings. The house in Figure 11B was drawn by a miss-classified HC subject who used less uniform strokes than those used by some patients. This can occur due to external factors such as education level, less contact with technology, or aging (the subject was 75 years old at the moment of the recording session). Conversely, drawings of PD patients detected as HC subjects (Figures 11A, and 11C) show relatively stable strokes, compared to those of the PD patient in Figure 11D, who is in severe state but was classified in intermediate state.

Table XIII indicates that the fusion of the three bio-signals improves the results in the classification of PD patients in different disease stages. Note that the fusion is highly accurate to detect HC subjects (100% with onset). Most of the miss-classified PD patients in low and severe stages are detected as PD patients in intermediate state. Note also that patients in severe stage are always miss-classified. We think that these

TABLE XIII
CLASSIFICATION OF HC SUBJECTS AND PD PATIENTS IN THREE STAGES OF THE DISEASE USING THE COMBINATION OF SPEECH, HANDWRITING, AND GAIT SIGNALS. $\mathbf{PD}_x$: PATIENTS IN LOW, INTERMEDIATE, AND SEVERE STATE ACCORDING TO THE MDS-UPDRS-III SCORE.

MDS-UPDRS-III score								
	onset UAR=55.6%				**offset UAR=45.2%**			
	HC	PD_1	PD_2	PD_3	HC	PD_1	PD_2	PD_3
HC	**100.0 (39)**	0.0 (0)	0.0 (0)	0.0 (0)	97.4 (38)	0.0 (0)	2.6 (1)	0.0 (1)
PD_1	0.0 (0)	**50.0 (7)**	42.9 (6)	7.1 (1)	21.4 (3)	**0.0 (0)**	78.6 (11)	0.0 (0)
PD_2	0.0 (0)	22.2 (4)	**72.2 (13)**	5.6 (1)	11.1 (2)	5.6 (1)	**83.3 (15)**	0.0 (0)
PD_3	0.0 (0)	66.7 (4)	33.3 (2)	**0.0 (0)**	16.7 (1)	0.0 (0)	83.3 (5)	**0.0 (0)**

results should improve with more data from patients in severe stage (we only had 6 PD patients in that stage). Regarding the classification of patients according to specific items of the scale for speech, upper limbs, and lower limbs, high UARs are obtained with handwriting and speech signals. The results obtained with speech signals to predict the speech item of the neurological scale are higher than those obtained when considering the total score. Confusion matrices show consistent results when predicting the total MDS-UPDRS-III scores. HC subjects are more accurately classified than patients in several stages of the disease. Patients in initial stages are commonly miss-classified as HC subjects and patients in severe stages are miss-classified in the intermediate stage.

VI. CONCLUSION

This paper presents a multimodal analysis of motor abilities of PD patients considering deep learning architectures based on TFRs and CNNs such that integrate information from speech, handwriting and gait signals. The proposed method models the difficulty of patients to start/stop the movement of muscles in lower and upper limbs, and in speech. Three main experiments were performed: (1) classification of PD patients and HC subjects, (2) classification of PD patients in different stages of the disease according to the total MDS-UPDS-III score, and (3) classification of PD patients in different stages of the disease according to specific impairments in lower and upper limbs, and in speech, considering sub-scores of the MDS-UPDRS-III scale. The experiments suggest that the proposed approach is highly accurate to classify PD patients and HC subjects using information of speech, handwriting, and gait separately. The results obtained with the proposed approach are higher than those obtained with traditional machine learning techniques. Additionally, the accuracy of the system improved up to 97.3% when information from the three bio-signals is merged. The classification of different stages of the disease shows that speech and handwriting are the most accurate. This fact can be explained because the transitions modeled in this study appear less frequently in gait than in speech or handwriting. It is necessary to evaluate other tasks with more transitions to obtain more accurate results. For instance, transitions that appear in the step cycle phase during the heel strike could show other gait impairments and increase the data to train more robust deep learning models. In order to do this, a more robust strategy to segment each step separately is needed to assess the onset/offset per step [41].

The models trained in this study show to be useful to characterize speech impairments of patients in three different languages: Spanish, German and Czech. This is validated only rapid repetitions of the syllables /pa-ta-ka/. Further experiments may be performed with more speech tasks to validate the language independence of the proposed approach.

The feature maps learned by the CNN trained with the multimodal data allow to interpret the hidden representations of the neural network. The first convolutional layers of the CNN trained with TFRs of speech show significant differences between PD patients and HC subjects. Similar results are obtained with the last layer of the CNN trained with handwriting.

The proposed approach seems to be promising to classify PD patients in different stages of the disease. The fusion of the three bio-signals is the most accurate approach to classify PD patients in different stages of the disease. The miss-classification errors appear mainly with patients in the initial stage which are miss-classified as HC subjects. Similarly, most of the patients in advanced stages are miss-classified as patients in intermediate stages of the disease, which indicates that the proposed approach makes errors that to some extent coincide with the natural progress of the disease.

CNNs seem to be suitable to model the difficulties of PD patients to start/stop the movements of different limbs, which allows the accurate classification of PD patients and HC subjects. In addition, the proposed architectures seem to be promising to classify different stages of the disease. Other architectures such as those based on recurrent neural networks and long short-term memory units should be considered in future works to model time-dependences of consecutive transitions and the co-articulation phenomena in speech. Recent advances in deep learning including the densely connected networks, or time-delay neural networks could be implemented as additional deep learning–based feature extraction approaches to model different bio-signals collected from PD patients.

The proposed approach can be extended to other applications also useful in the clinics. For instance it could be potentially used to detect prodromal stages of the disease, which would benefit the development of future neuroprotective therapies [42]. There is supporting evidence showing that the detection of prodromal stages of PD is possible from speech [9] and gait [43]. The main difficulty of these kinds of studies is to find the patients because it is necessary to recruit them before the disease to appear. Once the target group is found, it is required to start their monitoring over time in order to understand which are the patterns that become abnormal when early signs of the disease appear. Our research team in Medellín (Colombia) is currently collecting data from pre-clinical genetic subjects (people who have a gene mutation responsible for producing PD but with no clinical signs of the disease). We hope to find promising results in the near future. Another potential application for the proposed approach could be the discrimination between PD and other neurological disorders with similar symptoms, such as Huntington's disease or essential tremor. There is also evidence for this application in the literature, especially for speech [44] and gait [45].

ACKNOWLEDGMENT

This study was financed by CODI from University of Antioquia, grants PRG-2015–7683 and PRV16-2-01. This

project received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287. B. Eskofier gratefully acknowledges the support of the German Research Foundation (DFG) within the framework of the Heisenberg professorship program (Grant Number ES 434/8-1).

REFERENCES

[1] M. C. De Rijk, L. J. Launer, et al., "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts," *Neurology*, vol. 54, no. 5, pp. s214323, 2000.

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

[3] A. M. García, F. Carrillo, et al., "How language flows when movements don't: An automated analysis of spontaneous discourse in parkinson's disease," *Brain & Language*, vol. 162, no. Nov, pp. 19–28, 2016.

[4] C. G. Goetz et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[5] S. Patel et al., "Home monitoring of patients with Parkinson's disease via wearable technology and a web-based application," in *Int. Conf. of the IEEE Engineering in Medicine and Biology (EMBC)*, 2010, pp. 4411–4414.

[6] T Tykalova et al., "Distinct patterns of imprecise consonant articulation among parkinsons disease, progressive supranuclear palsy and multiple system atrophy," *Brain & language*, vol. 165, pp. 1–9, 2017.

[7] K. Forrest, G. Weismer, and G. S. Turner, "Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric adults," *Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2608–2622, 1989.

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

[9] J. Hlavnicka, R. Cmejla, 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.

[10] A. Tsanas et al., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

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

[12] M. Novotnỳ, J. Rusz, 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.

[13] S. Skodda et al., "Gender-related patterns of dysprosody in Parkinson disease and correlation between speech variables and motor symptoms," *J. of Voice*, vol. 25, no. 1, pp. 76–82, 2011.

[14] T. Grósz, R. Busa-Fekete, et al., "Assessing the degree of nativeness and Parkinsons condition using Gaussian processes and deep rectifier neural networks," in *Proceedings of INTERSPEECH*, 2015, pp. 1339–1343.

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

[16] B. Schuller, S. Steidl, et al., "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *Proceedings of INTERSPEECH*, 2015, pp. 478–482.

[17] M. Tu, V. Berisha, and J. Liss, "Interpretable objective assessment of dysarthric speech based on deep neural networks," in *Proceedings of INTERSPEECH*, 2017, pp. 1849–1853.

[18] M. Cernak, J.R. Orozco-Arroyave, F Rudzicz, H. Chirstensen, J.C. Vásquez-Correa, and Nöth E., "Characterisation of voice quality of Parkinsons disease using differential phonological posterior features," *Computer Speech & Language*, 2017.

[19] H. L. Teulings and G. E. Stelmach, "Control of stroke size, peak acceleration, and stroke duration in Parkinsonian handwriting," *Human Movement Science*, vol. 10, no. 2, pp. 315–334, 1991.

[20] O. Tucha, L. Mecklinger, et al., "Kinematic analysis of dopaminergic effects on skilled handwriting movements in Parkinson's disease," *Journal of neural transmission*, vol. 113, no. 5, pp. 609–623, 2006.

[21] H.L. Teulings and G.E. Stelmach, "Force amplitude and force duration in parkinsonian handwriting," *Tutorials in Motor Neuroscience*, vol. 62, pp. 149–160, 1991.

[22] J. Barth, M. Sünkel, et al., "Combined analysis of sensor data from hand and gait motor function improves automatic recognition of parkinson's disease," in *Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2012, pp. 5122–5125.

[23] P. Drotár, J. Mekyska, et al., "Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease," *Artificial intelligence in Medicine*, vol. 67, pp. 39–46, 2016.

[24] J. Klucken, J. Barth, et al., "Unbiased and mobile gait analysis detects motor impairment in Parkinson's disease," *PloS one*, vol. 8, no. 2, pp. e56956, 2013.

[25] F. Parisi et al., "Body-sensor-network-based kinematic characterization and comparative outlook of UPDRS scoring in leg agility, sit-to-stand, and gait tasks in Parkinson's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1777–1793, 2015.

[26] J. Hannink, H. Gaßner, et al., "Inertial sensor-based estimation of peak accelerations during heel-strike and loading as markers of impaired gait patterns in PD patients," *Basal Ganglia*, vol. 8, pp. 1, 2017.

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

[28] M Pastorino, MT Arredondo, J Cancela, and S Guillen, "Wearable sensor network for health monitoring: The case of Parkinson's disease," in *Journal of Physics: Conference Series*, 2013, vol. 450, p. 012055.

[29] J.C. Vásquez-Correa, J.R. Orozco-Arroyave, et al., "Multi-view representation learning via gcca for multimodal analysis of Parkinson's disease," in *Proceedings of ICASSP*, 2017, pp. 2966–2970.

[30] J. Rusz et al., "Effects of dopaminergic replacement therapy on motor speech disorders in parkinsons disease: longitudinal follow-up study on previously untreated patients," *Journal of Neural Transmission*, vol. 123, no. 4, pp. 379–387, 2016.

[31] K. Smulders et al., "Pharmacological treatment in parkinson's disease: Effects on gait," *Parkinsonism Relat. Risord.*, vol. 31, pp. 3–13, 2016.

[32] 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.

[33] J. Rusz et al., "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinsons disease," *Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[35] T. Tieleman and G. E. 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.

[36] F. Eyben et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[37] F. Eyben and B. Schuller, "opensmile:): The munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

[38] 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.

[39] G. J. Canter, "Speech characteristics of patients with Parkinsons disease: Iii. articulation, diadochokinesis, and over-all speech adequacy," *Journal of Speech and Hearing Disorders*, vol. 30, no. 3, pp. 217–224, 1965.

[40] S Pinto et al., "A cross-linguistic perspective to study of dysarthria in parkinsons disease," *Journal of Phonetics*, vol. 64, pp. 156–167, 2017.

[41] N. H. Ghassemi et al., "Segmentation of gait sequences in sensor-based movement analysis: A comparison of methods in parkinsons disease," *Sensors*, vol. 18, no. 1, pp. 145, 2018.

[42] R. B. Postuma et al., "Risk factors for neurodegeneration in idiopathic rapid eye movement sleep behavior disorder: a multicenter study," *Annals of neurology*, vol. 77, no. 5, pp. 830–839, 2015.

[43] L. Alibiglou et al., "Subliminal gait initiation deficits in rapid eye movement sleep behavior disorder: A harbinger of freezing of gait?," *Movement Disorders*, vol. 31, no. 11, pp. 1711–1719, 2016.

[44] J. Rusz et al., "Speech disorders reflect differing pathophysiology in parkinsons disease, progressive supranuclear palsy and multiple system atrophy," *Journal of neurology*, vol. 262, no. 4, pp. 992–1001, 2015.

[45] G. Ebersbach et al., "Clinical syndromes: Parkinsonian gait," *Movement Disorders*, vol. 28, no. 11, pp. 1552–1559, 2013.