Language-Independent Bimodal System for Early Parkinson's Disease Detection[⋆]

Catherine Taleb[1], Laurence Likforman-Sulem[2] and Chafic Mokbel[1]

[1] University of Balamand, Balamand El-Koura, Lebanon
[2] LTCI/Telecom Paris/Institut Polytechnique de Paris, Paris, France
catherine.taleb@std.balamand.edu.lb

Abstract. Parkinson's disease (PD) is a complex disorder characterized by several motor and non-motor symptoms that worsen over time, and that differ from person to another. In the early stages, when the symptoms are often incomplete, the diagnosis becomes difficult and at times, the subject may remain undiagnosed. This difficulty is a strong motivation for computer-based assessment tools that can aid in the early diagnosing and predicting the progression of PD. Handwriting's deterioration, vocal and eye movement impairments may be ones of the earliest indicators for the onset of the illness. A language independent model to handwriting, speech signals, and eye movement's recordings have been recently collected. After succeeding in building language independent models for PD early diagnosis using pure handwriting or speech, we propose in this work language independent models based on bimodal analyses (handwriting and speech), where both SVM and deep learning models are studied. Our experiments show that classification accuracy up to 100% can be obtained by our SVM model through handwriting/speech bimodal analysis.

Keywords: Parkinson's disease (PD), 2D CNN, 1D CNN-BLSTM, SVM, 1D CNN-MLP, handwriting, speech, data augmentation.

1 Introduction

PD is a neurological disorder caused by a decreased dopamine level on the brain. This disease is characterized by motor and non-motor symptoms that worsen over time. The motor symptoms consist of tremor, rigidity, slowness of movement or bradykinesia, micrographia, and speech difficulty [20]. In advanced stages of PD, clinical diagnosis is clear-cut. However, in the early stages, when the symptoms are often incomplete or subtle, the diagnosis becomes difficult and at times, the subject may remain undiagnosed. Furthermore, there are no efficient and reliable methods capable of achieving PD early diagnosis with certainty [2]. The difficulty in early detection is a strong motivation for computer-based assessment tools/decision support tools/test instruments that can aid in the early diagnosing and predicting the progression of PD [19]. Early detection of the

[⋆] Supported by organization x.

disease could be hugely beneficial in order for the patient to have access to a therapy that will slow down the course of PD progression. Handwriting's deterioration and vocal impairment may be ones of the earliest indicators for the onset of the illness [21], [22]. According to the reviewed literature, a language independent model to detect PD at early stages using multimodal signals has not been enough addressed. In our previous works [5], and [6], language-independent models for assessing the motor disorders in PD patients at early stages based on handwriting features have been developed; where two approaches were studied and compared: a classical feature extraction and classifier approach, and a deep learning approach. Also in [18] a language independent model based on pure speech analysis and SVM has been built. Approximately 97% classification accuracy was reached with both modalities and approaches. The main contribution of the present work is to build a language independent model for assessing the motor disorders in PD patients at early stages based on bimodal analysis (handwriting and speech), where two different approaches are studied and compared: handcrafted features and SVM, and deep learning.

The paper is organized as follows. In Section 2, we introduce our handwriting and speech datasets. In Section 3 an overview of related work is provided. Section 4 presents the bimodal system used for PD detection. We conducted several experiments that are described in Section 5. Conclusions and perspectives are drawn in Section 6.

2 Handwriting and Speech Datasets

Due to the lack of multimodal and multilingual PD database, a database (PD-MultiMC) that includes handwriting tasks, speech samples, and eye movements recordings has been collected from PD patients attending an experienced neurologist, in two phases ("on-state" (1 hour after taking L-dopa dosage for peak response to medication) and "off-state" (12 hours after the last L-dopa medication dosage)), and from Healthy control (HC) subjects selected from our entourage and seen by a neurologist to make sure they don't have any neurological disease. 21 PD patients (16 Males and 5 Females), and another 21 HC subjects (5 Males and 16 Females) are included in PDMultiMC database. PD and HC subjects are matching for age, years of education, and hand dominance. This database includes samples in three languages: 31 Arabic, 9 French, and 2 English and will be released on the IAPR TC11 repository. Even though the language representation is not balanced, language independence is somehow provided by averaging the signal (speech or handwriting). This average can be considered as the summation of a certain noise (summation of the average of channel information and the average of linguistic information) with disease characteristics; where the noise will not interfere in the classification. The modified Hoehn and Yahr (mH&Y) scale was measured to show the presence and the severity of PD motor symptoms. The mean mH&Y of PD patients in our database is 1.81±0.77; where around 95% of our PD patients show early to mild degree of disease severity. In this work, we focus on handwriting and speech modalities for early PD detection, where handwriting and speech samples are taken from HandPDMultiMC and SpeechPDMultiMC datasets (parts of PDMultiMC database). However, since our target is the early detection of the disease and since collecting large database at early stages is very difficult, and referring to [10] and [12], we make the assumption that mild stages can get nearer early stages one hour after taking L-dopa medication, since levodopa reduces the motor symptoms. For this reason, the seven handwriting tasks and the two speech tasks recorded for each of the 42 subjects in their "on-state" are studied and analyzed. Participants were asked to complete handwriting and speech tasks; where these tasks were chosen in a manner to highlight as much as possible the differences between PD and control, and where neurologists were consulted in the process of selecting the tasks.

Handwriting samples were collected using Wacom intuos 5 tablet with a sample rate of 197 points/s. The trace of the pen tip (X-Y-Z coordinate), the pressure of the pen tip on the surface, the angles of the pen relative to the tablet (altitude and azimuth), and timestamp were collected per sample point, forming seven times series as the handwriting dynamic signals. The seven handwriting tasks and the captured handwriting dynamic signals for a given task are displayed in Fig. 1. The first three tasks (with different degrees of complexity) demand one continuous pen movement, similar to spiral and meander tasks, which emphasis hypokinesia and tremor. Since there is no consensus on which stroke type reflects better the disease, we have included different type of strokes in the cursive tasks. From the other side, since PD patients have difficulty in maintaining constant force in long tasks, we decided to include words repetition in the other 4 tasks.

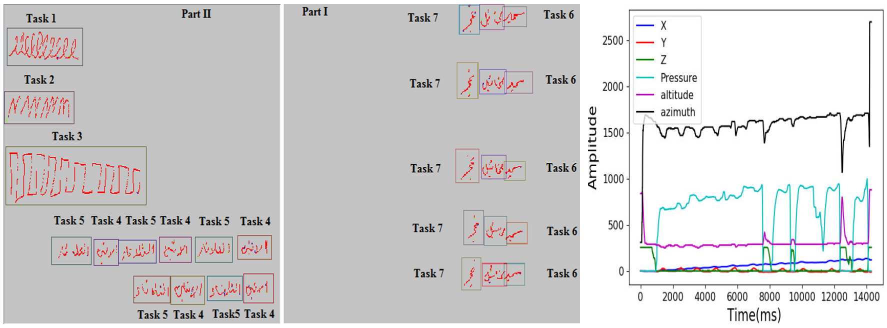

Fig. 1. The seven tasks segmented from the sheet filled by a PD subject and the handwriting dynamic signals captured for task1.

Speech wave signals were recorded using the internal microphone of the laptop (hp Elittbook 8570 w). Two channels sounds with 16-Bit depth and 44.1 KHz sampling rate were produced and saved. Two tasks were considered: 1) the participants were instructed to produce a sustained vowel 'a' (it is easy to be reproduced by elderly subjects and provides information about the phonatory and articulatory processes of speech production, and it was shown in [7] that the vowel ‘a’ reports the highest PD detection accuracy comparing to other vowels). 2) Participants were asked to read loudly a text appearing on the PC screen written with their familiar languages (such speech task permit judgments of speech rate, phrase length, voice quality, resonance, and precision of articulation, and also permits assessment of the prosodic features of speech [8]).

Many other complex tasks can more involve cognitive and functional issues and may better define the disease, but our main focus was on motor skills and the selected tasks can reflect most of the motor symptoms and can be easily reproduced by elderly people. Such tasks (except the sustained vowel) can also be used to assess dementia since they require linguistic skills, attention, and memory.

3 Single Modality PD Early Detection

3.1 Handwriting Modality

In literature, some works such as [4] focused on handwriting analysis for supporting early PD diagnosis where only patients with early to mild degrees of disease taken from PaHaW dataset [1] were studied. In our previous work [5] the main contribution was to find a feature selection approach for an improved PD early detection based on handwriting features suggested in [1]. Advanced language independent handwriting markers based on kinematic, stroke, pressure, entropy, and intrinsic features were extracted from the “on-paper” periods in each handwriting task (samples taken from HandPDMultiMC), forming a global feature vector of size 189. An SVM model was trained on these features, where a two-stage feature selection (FS) approach was applied. It was shown that handwriting can be a tool for PD early diagnosis with a 96.87% prediction accuracy when a set of kinematic, pressure, and correlation between kinematic and pressure features are used.

However, since hand-crafted features model required expert knowledge of the field, and since the database is small, pen-based features were learnt by means of deep learning where short term analysis is applied to avoid losing important information while applying global features extraction [6]. Two based learning models for end to end time series classification were proposed (the 2D CNN and the 1D CNN-BLSTM), where the whole handwriting dynamic signals have been studied so we can extract both in-air and on-surface features. Two new approaches were proposed to encode each handwriting dynamic signal into separate image for the 2D CNN model (spectrogram and modified Gramian Angular Field (GAF)), and compared to the approach proposed by Pereira et al. [3] that encode all the handwriting dynamic signals into single image. For the 1D CNN-BLSTM model, the raw signals are directly used. The number of handwriting dynamic signals k is a hyper-parameter varying between 1 and 7. Some pre-processing steps were applied: getting the same writing direction, and normalizing the X and Y coordinate. We have demonstrated the importance of both the 1D CNN-BLSTM, and the 2D CNN model with spectrograms as input in PD detection (Fig. 2)). These models have the ability to tackle the variation of information in time series either by explicitly considering the local short term information on the time axis of the non-stationary online handwriting signals or by dealing with raw time series directly. To cope with the limited data, and to improve our deep models, some data augmentation techniques (jittering, scaling, time-warping, and synthetic data generation) were applied on handwriting dynamic signals to generate new synthetic samples. We have found that combining both jittering and synthetic data augmentation techniques with the 1D CNN-BLSTM model yields 97.62% classification accuracy [6]. This diagnosis system has been validated on PaHaW database (closest to our database in term of handwriting dynamic signals compared to NewHandPD [3]), and it was found the importance of Z coordinates feature and the relevance of the results obtained (results are consistent on different datasets) [6].

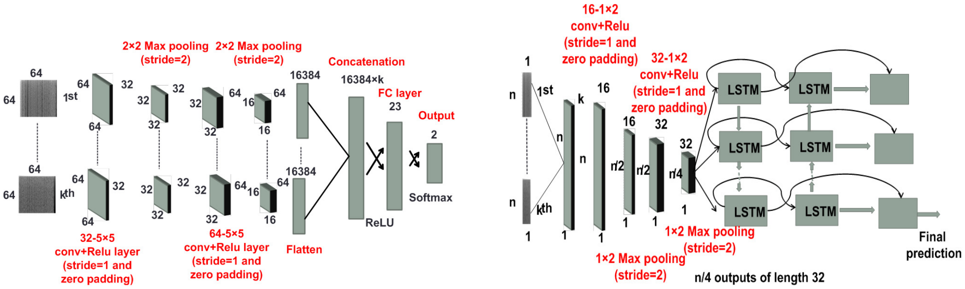
2×2 Max pooling (stride=2)
2×2 Max pooling (stride=2)
Concatenation
16384×k
FC layer
23
Output
2
16384
16384
64
64
1st
kth
32-5×5 conv+Relu layer (stride=1 and zero padding)
64-5×5 conv+Relu layer (stride=1 and zero padding)
Flatten
ReLU
Softmax
16-1×2 conv+Relu (stride=1 and zero padding)
32-1×2 conv+Relu (stride=1 and zero padding)
n
k
n/2
n/4
LSTM
1×2 Max pooling (stride=2)
1×2 Max pooling (stride=2)
n/4 outputs of length 32
Final prediction

Fig. 2. The left model refers to a single-task 2D CNN architecture. This architecture takes as input k two dimensional representations of 1D time series. The right model refers to a single-task 1D CNN-BLSTM architecture on a multivariate time series of length n. The output of the 1D CNN is fed to a BLSTM as a sequence of length n/4.

3.2 Speech Modality

Several studies in literature focused on the early detection of PD based on speech analysis such as Rusz et al. [14] who studied early staged PD patients with mean H&Y of 2.20±0.5. In [18], we have defined a language and task-independent acoustic feature set for assessing the motor disorders in PD patients, and have studied the influence of sampling rate and unvoiced sounds on the performance. Only phonation and articulation handcrafted features are studied, where the prosody features are excluded since they depend on the language spoken [9]. Some pre-processing steps were applied prior extracting the low level descriptors (LLD) features: removing silence at the start and the end of the speech, removing speech that does not refer to the subject and each spontaneous intervention introduced by the subject that was not directly related with the task, converting the 2 channels signal into mono signal, and down-sampling signal rate. The LLDs were extracted over 20ms frames shifted by 10ms from the processed voice signal using openSMILE toolkit. Global features were obtained from the z-scored

LLD features by applying some statistical functions. This resulted in 220 global features per task. An SVM model was trained on these features, where the two-stage feature selection approach proposed in [5] was applied. Unvoiced sounds and sampling rate effects on classification performance of PD detection through voice analysis were studied. Our language independent SVM model for PD early diagnosis through voice analysis achieved 97.62% accuracy. It was found that the effect of sampling rate on PD classification may depend on task and features used. We have found that signals with low sampling rate (less than 16 KHz) can lose valuable information that can play a good role in PD detection, where a sampling rate of 24 KHz for sustained vowel 'a' and text reading (voiced sounds) tasks, and 16 KHz for text reading task (voiced and unvoiced sounds) are appropriate for the features analyzed. We have also demonstrated the importance of unvoiced frames in PD detection and the importance of MFCC coefficients to quantify the problems in speech articulation and to detect the disease [11].

4 Bimodal PD Early Detection

The main contribution of this work is to build a language independent bimodal system for assessing the motor disorders in PD patients at early stages based on handwriting and speech signals. Since there is no consensus on which modality is more appropriate to help on PD diagnosis in early stages; so combining and analyzing both signals may deliver a more accurate PD prediction. Two different learning approaches were applied: feature-based and deep learning approaches. Fusion of different modalities can be executed at different levels: data level, short term or global features level, or decision level as shown in Fig. 3. In this work, only global features and decision level fusion are applied since data-level and short term feature-level fusions are used when the multiple raw data even come from a same type of modality source, or are synchronized (which is not our case) [13].

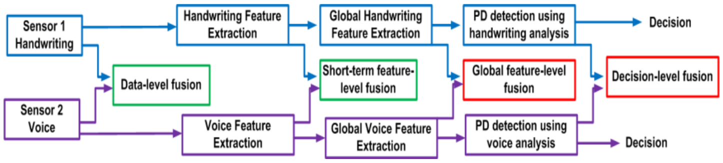
Sensor 1 Handwriting
Handwriting Feature Extraction
Global Handwriting Feature Extraction
PD detection using handwriting analysis
Decision
Data-level fusion
Short-term feature-level fusion
Global feature-level fusion
Decision-level fusion
Sensor 2 Voice
Voice Feature Extraction
Global Voice Feature Extraction
PD detection using voice analysis
Decision

Fig. 3. Levels of bimodal fusion of handwriting with speech.

4.1 Feature-Based Approach

In this section, only global feature-level fusion is applied. Global feature-level fusion consists in combining two global features vectors, one for each modality. Each vector includes information of all tasks per subject. For each modality, the set of global and language-independent features defined in Section 3 (189 "on-paper" global handwriting features and the 220 global acoustic features) are extracted for each task then combined together to form a single feature vector. Two different methods are applied to combine the feature vectors in each modality: calculate the average feature vector across the different tasks, or concatenate the features vectors together. At a later stage, the features vectors obtained from the 2 modalities will be concatenated to form a global bimodal vector that will be used to detect PD using a SVM model with RBF kernel; where feature-level Min-Max is applied on each feature separately before classification. An overview of the SVM model trained on bimodal pre-engineered features is shown in Fig. 4.

For speech, all the pre-processing steps described in Section 3 are applied to the voice signal before extracting the LLD features, and speaker level z-normalization is applied to the LLD features to reduce the effects of variations that are not related to the disease (such as recording environment noise, speaking styles or accent etc.). Based on the results obtained in [18], the sustained vowel 'a' is sampled at 24 KHz while the text reading is sampled at 16 KHz and both voiced and unvoiced frames in the text reading are studied. The two-stage feature selection approach defined in [5] is also applied here, where the first stage consists of a pure statistical analysis of the data (where a sequence of significance levels between 0 and 1 were tested and the one with the best validation accuracy will be picked) and the second stage consists of applying a suboptimal approach that provides a kind of benchmark of the relevance of the features in the desired task.

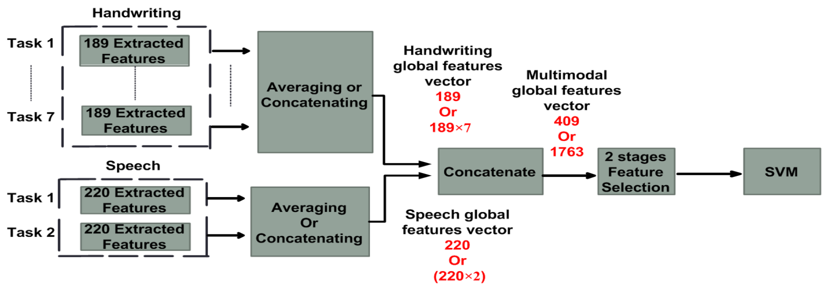
Handwriting
Task 1
189 Extracted Features
Task 7
189 Extracted Features
Averaging or Concatenating
Handwriting global features vector
189
Or
189×7
Multimodal global features vector
409
Or
1763
Speech
Task 1
220 Extracted Features
Task 2
220 Extracted Features
Averaging Or Concatenating
Concatenate
2 stages Feature Selection
SVM
Speech global features vector
220
Or
(220×2)

Fig. 4. Levels of bimodal fusion of handwriting with speechSVM trained on bimodal pre-engineered features using global feature-level fusion.

4.2 Deep-Learning Approach

The core idea is to train our model with language-independent bimodal feature vector. For deep learning approach, to obtain language-independent feature vector, the model is trained on all the languages so the features will not be biased toward a specific language. Handwriting and audio pre-processing steps mentioned in Section 3 are applied in this section. For text reading task, the best sounds combination found in the first part of this work are also applied here.

However, the voice sampling rate for both tasks is set to 8 KHz to reduce computational time, and memory usage. For all the deep models, the number of hidden nodes is selected in a way to have a number of independent parameters smaller than the number of data points available.

2D CNN/2D CNN. One of the deep learning models studied is the 2D CNN model with spectrogram images (found in [6] and summarized in Fig. 2), which is applied in this work in both modalities; where spectrogram 2D representation for both raw handwriting signals and speech signals are obtained by applying Short Term Fourier Transform (STFT). Blackman windowing function is applied, where both the window length and the number of Fast Fourier Transform (FFT) points are set to 256 and the overlapping rate is 50%. Lanczos technique [18] is used to ensure that the number of input feature maps is identical for all subjects by resizing the spectrogram images to 64×64 pixels resolution. This model can be used for classification from a single image (voice case), or classification from k measurements, where each measurement is encoded into spectrogram image (handwriting case). The number of handwriting dynamic signals k is a hyper-parameter varying between 1 and 7. Fusion of both handwriting and voice modalities are executed here at two levels: global feature-level and decision-level.

Global feature-level fusion. The aim here is to form one feature vector with information of all tasks per subject and per bio-signal. To do this, for each modality individual 2D CNNs are trained per task and the feature maps obtained by the convolutional layers (the output of the concatenated layer in Fig. 2) for each task are combined together whether by averaging (model M1) or by concatenating (model M2). The embeddings obtained from the 2 modalities are concatenated to form a bimodal vector per subject. The created feature vectors are then used to classify PD patients and HC subjects using fully connected layers as shown in Fig. 5.

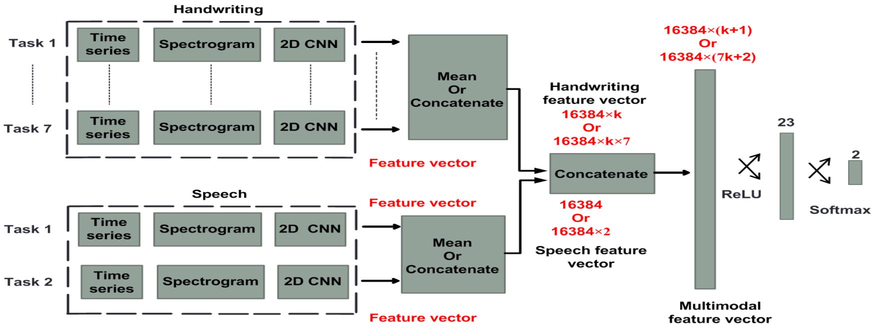

Fig. 5. Bimodal assessment using 2D CNN models and global feature-level fusion.

Decision-level fusion. For voice modality, two cases are studied: the first case (defined by case 1 and model M3) is when spectrogram is obtained for the whole audio signal (image input of size 64×64 pixels and 23 hidden nodes), and the second case (defined by case 2 and model M4) is when the audio signal is cropped into segments of size 4s in order to increase the number of samples, and to keep the nonlinear variation over the time axis as shown in Fig. 6. Spectrogram is obtained for each segment separately (image input of size 249×129 pixels and 1024 hidden nodes). All the training segment slices images are considered independent training instances. Segmentation is also applied when predicting the label of a testing time series. No window slices referring to the same participant exist in training and test. To make the final prediction for each subject in the test set, the S probability vectors outputs of the 2D CNN models are considered as a Multivariate sequence of length S, and are used as input to a dynamic BLSTM to decide the final prediction. Two multilayer perceptron (MLPs) models (MLP1 and MLP2) are applied, where each one is used to combine the probability vectors (each of size 2) obtained by all tasks in each modality. At a later stage, another MLP model (MLP3) is used to combine the probability vectors provided by each of MLP1 and MLP2 (each of size 2) in order to get the final prediction (refer to Fig. 6).

1D CNN-BLSTM/1D CNN-MLP. The 1D CNN-BLSTM model with raw time series as input (see Fig. 2) is applied in this section. However, since we are working with long audio signals, we have decided to crop the audio signal into segments of fix length (4s) and apply the 1D CNN-MLP model (shown in Fig. 7). This model is defined by M5. The number of handwriting dynamic signals k is a hyper-parameter varying between 1 and 7. Fusion of both handwriting and voice modalities are executed here at decision-level only since the feature maps in both modalities differ (the features map in handwriting modality is a sequence of length n/4 of vectors of size 32, where in voice modality the features map is a vector of length 32×n/4), in addition in each modality n varies from one task to another and the number of audio segments of length 4s varies between tasks. Individual 1D CNN-BLSTMs are trained for each task in handwriting modality, where individual 1D CNN-MLPs followed by BLSTMs to make the final prediction are trained for each task in voice modality as shown in Fig. 7. Three different MLPs are also applied to get the final prediction.

1D CNN-BLSTM/2D CNN. The 1D CNN-BLSTMs and 2D CNNs models are combined. For voice, 2D CNNs are applied with the STFT as input, where for handwriting 1D CNN-BLSTMs are used with the raw signals as input. Fusion of both handwriting and voice modalities are also executed here at decision-level, and the number of handwriting dynamic signals k is a hyper-parameter varying between 1 and 7. For voice modality, spectrogram is even obtained for the whole audio signal (defined by model M6), or for each 4s segment as described previously (defined by M7). Individual 1D CNN-BLSTMs and 2D CNNs are trained for each task in handwriting and voice modalities respectively, where

MLPs are also applied to combine and obtain the final prediction (similar to Fig. 6, where the only difference is that for handwriting 1D CNN-BLSTMs are applied with the raw signals).

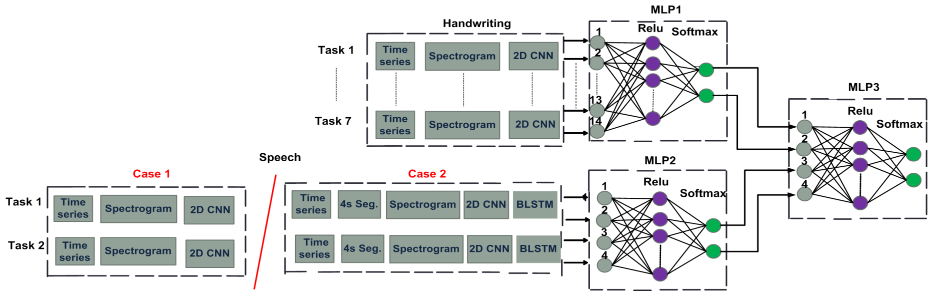

Fig. 6. Bimodal assessment using 2D CNN models and decision-level fusion.

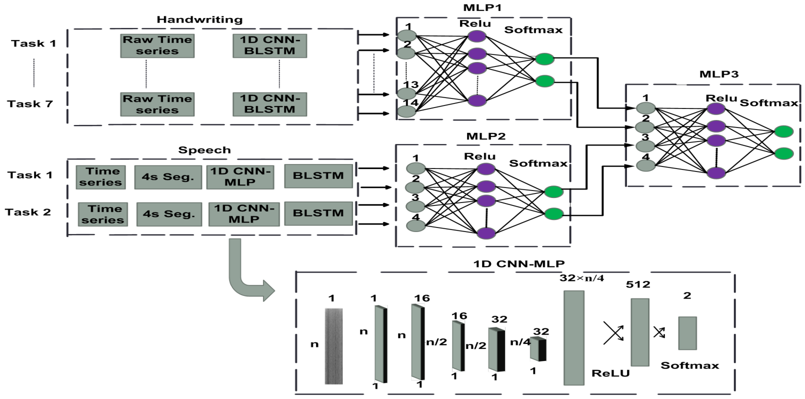

Fig. 7. Bimodal assessment using 1D CNN-BLSTM and 1D CNN-MLP models and decision-level fusion.

4.3 Data Augmentation

Based on the findings in [6], data augmentation applied to time series improves the 1D CNN-BLSTM model performance, and fails to improve the 2D CNN model performance. In addition, we have found the power of combining both jittering and synthetic data augmentation techniques with the 1D CNN-BLSTM model. Based on these findings, and after selecting the best bimodal systems, these techniques are applied. However, synthetic data generation is memory consuming method as long as we are working with long audio signals. For this reason, only jittering data augmentation method is applied with voice modality; where in handwriting modality, jittering is applied with 2D CNNs and the combination of jittering and synthetic data is applied with the 1D CNN-BLSTMs. For jittering, several values of noise intensity are studied in order to explore its effect on classification. In [6] we have found that the best results are obtained when the training data is augmented twice. This has also been adopted in this work.

5 Experiments and Results

In this work, bimodal assessment of PD is studied where both SVM model trained on handcrafted features and deep models are studied and compared, where the 42 subjects are divided into 3 folds, with the 66.66/33.33% (training/validation) proportion using stratified sampling method. Sequentially, one fold is validated using the classifier trained on the remaining 2 folds. The total accuracy is obtained by calculating the mean of all the folds accuracies. We have decided not to use a separate test set due to a low database size. As a result the validation set can be considered as test set. Starting with the SVM model based on global feature-level fusion defined in Section 4.1. The significance level with the best validation accuracy, the number of selected features and the performances obtained with one and two stage feature selection methods are shown in Table 1.

Table 1. Table of comparison between the performance obtained with one and two stage feature selection methods.

	Method	**1 stage FS**	**2 stages FS**
Average	Acc (Sens,Spec) (%)	92.86 (95.24, 90.48)	100 (100, 100)
	Significance level	0.0933	
	# Selected features	125	22
Concatenation	Acc (Sens,Spec) (%)	97.62 (95.24, 100)	100 (100, 100)
	Significance level	0.0350	
	# Selected features	356	55

The highest classification accuracy obtained with average method is up to 100% for N=22 features (15 handwriting and 7 acoustic). For concatenation method, also the highest classification accuracy obtained is up to 100% for N=55 features (52 handwriting and 3 acoustic). Most of the selected features providing the best performance for both methods include kinematic, pressure, and correlation between kinematic and pressure features for handwriting modality, and MFCC coefficients for voice modality; agreed with the conclusion found in [5] and [18]. From a clinical point of view, acceleration and stroke size are regulated by the motor control of wrist and finger movement (mechanism inaccurate in PD). Moreover, pressure features can give further detailed information that cannot be obtained from kinematic features, hence, the significance to show the relationship between kinematic and pressure features [5]. From the other side PD affects the movement of the articulatory muscles resulting varying energy in frequency bands of speech signal, and MFCCs coefficients can compute these variations [18]. This can explain the frequent existence ofsuch features in the selected set.

Moving to the bimodal assessment using deep learning, the three different combinations described in Section 4.2 were studied and compared. The results obtained are summarized in Table 2. For the 2D CNN model, decision-level fusion method performs better than feature-level fusion. This can be related to the fact that feature-level fusion is effective when time synchronized modalities are to be fused (fusion of speech and eye movements for example) [13]. The best models are selected and presented in Table 3 ; where the 'All-tasks' performances of both modalities beside the bimodal system are shown. Once the best models are selected, data augmentation techniques described in Section 4.3 are applied, where for jittering a random additive scalar is sampled from a Gaussian distribution with 0.3 STD for handwriting modality and 0.1 STD for voice modality. The new time series are either used directly with the 1D CNN-BLSTM or converted into spectrogram images with the 2D CNN. The 'All-tasks' performances of both modalities beside the bimodal system are shown in Table 4, and Task-wise accuracies in Table 5 respectively. Based on these results, we can see that the best handwriting features combination found is the same as the one found with handwriting modality [6], and it is clear how deep learned audio features has no effect on PD detection, and how the results obtained when applying deep learning to detect PD from raw speech signals are not satisfactory as the ones obtained with handwriting analysis. It is challenging to learn acoustic deep models from raw signals and especially when very few convolutional layers are used for acoustic feature extraction, which might be insufficient for building high-level discriminative features [15]. In our case, since we are working with small dataset, it will not be a good idea to enlarge our model. We investigate that it is more efficient to build the model using low-level audio descriptors instead of applying the raw audio waveforms directly.

Table 2. Bimodal classification of PD and control performance, where the seven models are defined in Section 4.

Model	**Fusion**	**Modality**	**Best handwriting**	**Bimodal**
	method	**combination**	**Features combination**	**Acc (Sens, Spec) %**
M1	Feature-level	Averaging	Z	78.57 (85.71,71.43)
M2		Concatenating	X+Y+Z+ Pre.+ Alt.+Azi.	61.9 (61.9, 61.9)
M3	Decision-level	MLP	X+Y+Z+ Pre.+ Alt.	**83.33** (87.71, 80.95)
M4			X+Y+Z+ Pre.+ Alt.	**83.33** (87.71, 80.95)
M5			X+Y+Z+ Pre.+ Alt.+Azi.	**88.1** (80.95, 95.24)
M6			X+Y+Z+ Pre.+ Alt.+Azi.	**88.1** (80.95, 95.24)
M7			X+Y+Z+ Pre.+ Alt.+Azi.	**88.1** (80.95, 95.24)

The log-spectrogram offers a rich representation of the temporal and spectral structure of the input signal. The use of log-spectrograms is thus studied. According to Table 4, the results show how the accuracy performance is improved from 52.38% to 71.43% after considering the log-spectrogram as input instead of the raw signal. However, the achieved results are still non-satisfactory compared to the results obtained with handwriting. A possible explanation of this behavior is that in spectrogram representations, it is difficult to separate simultaneous sounds since they all sum together into a distinct whole [16]. This means that a particular observed frequency in a spectrogram cannot be assumed to belong to a single sound. In addition, moving a sound vertically in a spectrogram might influence the meaning. Therefore, the spatial invariance that 2D CNNs provide might not perform as well for this form of data [16]. Finally, periodic sounds comprised of a fundamental frequency and a number of harmonics which are most often non-locally distributed on the spectrogram. Finding local features in spectrograms using 2D convolutions will be complicated in this case [16].

Table 3. Handwriting, audio and bimodal classification performance (Acc (Sen, Spec)) obtained with decision-level fusion method.

Model	Hand. Perf. (%)	Voice Perf. (%)	Bimodal Perf. (%)
M3	83.33 (87.71, 80.95)	54.76 (76.19, 33.33)	83.33 (87.71, 80.95)
M4	83.33 (87.71, 80.95)	52.38 (61.9, 42.86)	83.33 (87.71, 80.95)
M5	88.1 (80.95, 95.24)	54.76 (23.81, 85.71)	**88.1** (80.95, 95.24)
M6	88.1 (80.95, 95.24)	54.76 (76.19, 33.33)	**88.1** (80.95, 95.24)
M7	88.1 (80.95, 95.24)	52.38 (61.9, 42.86)	**88.1** (80.95, 95.24)

Table 4. Performance measures obtained after applying data augmentation and decision-level fusion, where the best handwriting features combination are the ones found in Table 2.

Model	Aug. Technique	Hand. Perf. (%)	Voice Perf. (%)	Bimoda Perf. (%)
	Hand/Voice	Acc (Sens, Spec)	Acc (Sens, Spec)	Acc (Sens, Spec)
M3	Jitter/Jitter	83.33(87.71,80.95)	57.14(95.24,19.05)	83.33(87.71,80.95)
M4	Jitter/Jitter	83.33(87.71,80.95)	71.43(76.13,66.67)	85.71 (71.43, 100)
M5	Jitter+Syn./Jitter	**97.62**(95.24,100)	52.38(33.33,71.43)	**97.62** (95.24, 100)
M6	Jitter+ Syn./Jitter	**97.62**(95.24,100)	57.14(95.24,19.05)	**97.62** (95.24, 100)
M7	Jitter+ Syn./Jitter	**97.62**(95.24,100)	71.43(76.13,66.67)	95.24(95.24,95.24)

Nevertheless, cropping the audio signal into short segments and getting the log-spectrograms of each segment (short-term analysis) seems to be more effective than getting the log-spectrograms of the whole signal (global analysis) with 2D CNN (accuracy from 57.14% to 71.43%). We believe that this is due to the larger number of samples needed to train the 2D CNN, and to the idea of maintaining the nonlinear variation over the time axis. Data augmentation improves the 2D CNN model performance when audio signal segmentation is applied, and fails to improve the 1D CNN-MLP with raw signal and 2D CNN without segmentation performance. In [6], we have found that data augmentation does not improve the 2D CNN model with online handwriting spectrograms since the augmented time series are converted into spectrograms then normalized to a fixed dimension. While here, since no normalization is applied on spectrograms, this means that the model may benefit the most from the new generated signals.

Finally, from a quick analysis of the Task-wise accuracies presented in Table 5, text reading task performs better than the sustained phonation vowel ‘a’ in PD detection. We believe that the text reading task is richer in terms of acoustic and prosodic information, which makes them more convenient for automatic PD detection in contrast to maximum phonation time of vowel ‘a’ which contains less information [17]. The same conclusion was found in [18] when a SVM model was trained on pure handcrafted acoustic features.

In our opinion, it may be more efficient to build a deep model using some low level acoustic descriptors instead of using speech signal. From the other side, the acoustic handcrafted features proposed in this work have achieved good results in pure speech and in bimodal corpuses. In general, deep learning models are basically selected to avoid handcrafted features extraction that needs an expert knowledge of the field, or to achieve a better result by extracting deep features.

Table 5. Task-wise system and all-tasks system accuracies (in %) for various models and training schemes presented in Table 4.

Task	**M3**	**M4**	**M5**	**M6**	**M7**
Aug. technique	Jitter	Jitter	Jitter/Syn.	Jitter/Syn.	Jitter/Syn.
Repetitive letter ‘l’	**69.05**	**69.05**	59.52/47.62	59.52/47.62	59.52/47.62
Triangular wave	71.43	71.43	**80.95**/78.57	**80.95**/78.57	**80.95**/78.57
Rectangular wave	61.9	61.9	71.43/**76.19**	71.43/**76.19**	71.43/**76.19**
Repetitive “Monday”	59.52	59.52	**78.57**/76.19	**78.57**/76.19	**78.57**/76.19
Repetitive “Tuesday”	**71.43**	**71.43**	57.14/59.52	57.14/59.52	57.14/59.52
Repetitive “Name”	52.38	52.38	**57.14**/50	**57.14**/50	**57.14**/50
Repetitive “Family Name”	**71.43**	**71.43**	69.05/64.29	69.05/64.29	69.05/64.29
Aug. technique	Jitter	Jitter	Jitter	Jitter	Jitter
Sustained vowel ‘a’	52.38	**66.67**	52.38	52.38	**66.67**
Text reading	54.76	**73.81**	54.76	54.76	**73.81**
All tasks	83.33	85.71	**97.62**	**97.62**	95.24

6 Conclusions

The main contribution of this work is to build a language independent bimodal system for PD early diagnosis by combining both handwriting and speech signals, where this combination may be more appropriate to help on PD diagnosis in early stages. Both SVM and deep learning models are studied and compared in this work. For text reading task, the best combination (only voiced or the combination of voiced and unvoiced sounds) found in our previous work [18] is applied in this work with both approaches (SVM and deep learning). However, the best sampling rates found in [18] are only applied with the SVM, where a low sampling rate value (8 KHz) was applied with deep models (for memory usage problem). The results obtained with the SVM model are better than the ones obtained with deep learning. We have found how SVM with the combination of information from both handwriting and speech modalities deliver a more accurate PD prediction (accuracy up to 100% that needs to be confirmed on larger scaled data) than pure handwriting (96.87% accuracy [5]) and speech (97.62% accuracy [18]) analyses.

The observations and conclusions obtained in this work are many. We have found that decision-level fusion method performs better than feature-level fusion in case of combining non-synchronized signals. In addition we have noticed how it is challenging to learn acoustic deep models from raw signals and especially when very few convolutional layers are used for acoustic feature extraction. Feeding a CNN with 2D spectrograms performs better than the raw signals but the results are still non-satisfactory. Higher number of training data samples and preservation of the signal nonlinear variation over the time axis improve the performance. Text reading task performs better than the sustained phonation vowel 'a' due to the existence of acoustic and prosodic information. Data augmentation methods applied on voice signals may increase deep learning model performance when the raw signals are converted into 2D spectrograms and the non-linearity over time axis is preserved. Deep models with the combination of handwriting and speech modalities deliver same PD prediction accuracy as pure handwriting analysis, and more accurate PD prediction than pure speech analysis. Since we believe that it may be more efficient to build deep models with some low level acoustic descriptors as inputs instead of raw speech signals in case we are working with small database, as a future work it will important to build such model to approve the effectiveness of speech analysis in PD detection. Despite the encouraging results obtained, there are still some works to do before putting our PD detection bimodal model into clinical use due to the fact that we have few subjects, in comparison with the real world where we would have thousands of subjects, but the findings in this work can form a solid basis to a future stage of research that needs to involve a much larger set of patients. For this reason, the observations and conclusions obtained in this work and the relevance of our system should be validated on a larger scaled database in future work (since for the time being we can not find any another public bimodal database (handwriting and speech) that can be used to validate our diagnosis system).

References

1. Drotar, P., Mekyska, J., Rektorova, I., Masarova, L., Smekal, Z., Faundez-Zanuy, M.: Decision support framework for Parkinsons disease based on novel handwriting markers. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 1-1 (2015). https://doi.org/10.1109/tnsre.2014.2359997
2. Weiner, W.J., Shulman, L.M., Lang, A.E.: Parkinsons disease: a complete guide for patients and families. Johns Hopkins Univ Pr (2013)
3. Pereira, C.R., et al.: Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification. Artificial Intelligence in Medicine. **87**, 67-77 (2018)
4. Impedovo, D., Pirlo, G., Vessio, G.: Dynamic Handwriting Analysis for Supporting Earlier Parkinson's Disease Diagnosis. Information. **9**, 247 (2018)
5. Taleb, C., Likforman-Sulem, L., Khachab, M., Mokbel, C: Feature Selection for an Improved Parkinson's Disease Identification Based on Handwriting. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France (2017)

6. Taleb, C., Likforman-Sulem, L., Mokbel, C., Khachab, M.: Detection of Parkinson's disease from handwriting using deep learning: A comparative study. Evolutionary Intelligence. (2020). https://doi.org/10.1007/s12065-020-00470-0
7. Orozco-Arroyave, J.R., et al.: Characterization Methods for the Detection of Multiple Voice Disorders: Neurological, Functional, and Laryngeal Diseases. IEEE Journal of Biomedical and Health Informatics. **19**, 1820–1828 (2015)
8. Duffy, J.: Motor Speech Disorders: Clues to Neurologic Diagnosis. Parkinson's Disease and Movement Disorders: Diagnosis and Treatment Guidelines for the Practicing Physician. 35–53 (2000). https://doi.org/10.1016/j.wocn.2017.01.009
9. Pinto, S., Chan, A., Guimarães, I., Rothe-Neves, R., Sadat, J.: A cross-linguistic perspective to the study of dysarthria in Parkinson's disease. Journal of Phonetics. **64**, 156–167 (2017)
10. Shalash, A.S., et al.: Non-Motor Symptoms as Predictors of Quality of Life in Egyptian Patients with Parkinson's Disease: A Cross-Sectional Study Using a Culturally Adapted 39-Item Parkinson's Disease Questionnaire. Frontiers in Neurology. **9**, (2018). https://doi.org/10.3389/fneur.2018.00357
11. Khan, T.: Running-speech MFCC are better markers of Parkinsonian speech deficits than vowel phonation and diadochokinetic (2014). http://www.diva-portal.org/smash/record.jsf?pid=diva2:705196. Last accessed 10 May 2021
12. Mazuel, L., et al: Proton MR Spectroscopy for Diagnosis and Evaluation of Treatment Efficacy in Parkinson Disease. Radiology. **278**, 505-513 (2016)
13. Dumas, B., Lalanne, D., Oviatt, S.: Multimodal Interfaces: A Survey of Principles, Models and Frameworks. Lecture Notes in Computer Science Human Machine Interaction. 3-26 (2009)
14. Rusz, J., et al.: Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. The Journal of the Acoustical Society of America. **134**, 2171–2181 (2013). https://doi.org/10.1121/1.4816541
15. Dai, W., Dai, C., Qu, S., Li, J., Das, S.: Very deep convolutional neural networks for raw waveforms. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA (2017)
16. Lonce, W.: Audio spectrogram representations for processing with convolutional neural networks. In: Proceedings of the 1st International Workshop on Deep Learning for Music, Anchorage, AK, USA (2017)
17. Pompili, A., et al.: Automatic Detection of Parkinson's Disease: An Experimental Analysis of Common Speech Production Tasks Used for Diagnosis. Text, Speech, and Dialogue Lecture Notes in Computer Science. 411-419 (2017)
18. Taleb, C.: Parkinson's disease detection by multimodal analysis combining handwriting and speech signals (Unpublished doctoral dissertation). Telecom Paris, France (2020)
19. Jeancolas, L., et al.: Automatic detection of early stages of Parkinsons disease through acoustic voice analysis with mel-frequency cepstral coefficients. 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). (2017). https://doi.org/10.1109/atsip.2017.8075567
20. Sharma, R.K., Gupta, A.K.: Voice Analysis for Telediagnosis of Parkinson Disease Using Artificial Neural Networks and Support Vector Machines. International Journal of Intelligent Systems and Applications. **7**, 41-47 (2015)
21. Little, M., Mcsharry, P., Hunter, E., Spielman, J., Ramig, L.: Suitability of Dysphonia Measurements for Telemonitoring of Parkinsons Disease. IEEE Transactions on Biomedical Engineering.**56**, 1015–1022 (2009)
22. Rosenblum, S., Samuel, M., Zlotnik, I., Schlesinger, I.: Handwriting as an objective tool for Parkinson's disease diagnosis. Journal of Neurology. **260**, 2357–2361 (2013)