THIS IS AN EARLIER VERSION OF THE PAPER WITH DOI:
10.1016/j.bspc.2022.104281

Addressing smartphone mismatch in Parkinson's disease detection aid systems based on speech

Mario Madruga[a,*], Yolanda Campos-Roca[b], Carlos J. Pérez[a]

[a] *Universidad de Extremadura, Departamento de Matemáticas, Spain*
[b] *Universidad de Extremadura, Departamento de Tecnología de los Computadores y las Comunicaciones, Spain*

Abstract

Objective: Voice analysis based systems offer low-cost, highly available automatic diagnostic aid for Parkinson's disease (PD) detection anywhere a smartphone with a broadband connection is available. However, reliability depends on factors affecting the communication channel. In this paper the effects of recording device mismatch are analyzed. Multicondition training (MCT) is proposed to improve robustness against that mismatch. **Methods:** An experiment on 30 PD patients and 30 healthy subjects was designed. 3 vocalizations of sustained \a\ were recorded using a smartphone. These recordings, along with a simulation of 8 additional smartphones, were analyzed. Acoustical features were extracted and averaged per patient and recording device. Machine learning was used to distinguish healthy from PD patients by using different combinations of train-test smartphones. **Results:** By using the same device for training and testing, a 10% best-worse mean accuracy drop is observed. The gap among different devices reaches 37%. MCT retains 90% of the maximum accuracy and exceeds a 20% mean accuracy while lowers dispersion of the aggregated results obtained with single condition. Smartphone position shows a direct impact on performance. **Conclusion:** Recording device has a major effect on results. It is also found that red positioning of the recording device might also be influential. Using MCT appears to improve robustness. **Significance:** Results support the use of mobile devices to create an automated PD detection test. It is also encouraged to consider the use of MCT to obtain more robust and reliable results across different devices.

Keywords: Parkinson's disease, Microphone simulation, Machine learning, Diagnosis aid, Channel mismatch robustness.

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disorder usually classified as a motor function disease. It is characterized by the presence of bradykinesia, rigidity and tremor [1]. It is estimated that more than 8 million people worldwide suffer from PD. The population group aged over 65 accumulates most of

*Corresponding author
Email addresses: mariome@unex.es (Mario Madruga), ycampos@unex.es (Yolanda Campos-Roca), carper@unex.es (Carlos J. Pérez)

the patients, and the percentage rapidly grows as the population reaches 80 years old. The prevalence shows an age-standardized rate of 106.28 per 100,000 inhabitants, with an increasing percentage change of 155.5% in the 1990-2019 period [2].

A reliable diagnostic test for PD has yet to be available. A range of novel techniques have been developed in order to obtain an early PD diagnosis. Examples are found in [3], using electroencephalograms; [4] finds markers in magnetic resonance images; or [5], analyzing motion of upper and lower extremities. However, their availability as a general diagnostic method is low.

Voice analysis has been proposed as a non-invasive low-cost method for PD detection and assessment. 75%-95% of people with PD suffers from some sort of speech impairment [6], so voice analysis is a potential candidate to become an additional biomarker that can be used for PD diagnosis. This has led to the research of early detection of PD by analyzing different aspects of voice impairment, for which sustained vowels [7], running speech [8], and diadochokinesis tests [9] have been considered. Also, a variety of machine learning techniques have been proposed, from classical approaches [10] to state-of-the-art deep learning methods [11]. [12, 13] provide a thorough review of voice assessment approaches in the context of PD and other diseases.

One of the advantages that these non-invasive diagnostic techniques offer is ubiquity. The omnipresence of mobile technology allows to carry a recording device with broadband connectivity in the form of a smartphone. This technology gives both practitioners and patients access to advanced diagnostic aid tools almost everywhere. In fact, PD related telemedicine systems have long been developed [14], with a recent focus on mobile devices [15, 16].

The concept of channel robustness is commonly applied in relation to speech classification systems, meaning that perturbations affecting the channel do not critically decrease the system performance. The use of the term robustness with this meaning is often present in the scientific literature related to speech classification systems [17].

Channel robustness covers several factors that may produce variations in the outcomes of the experiments (noise, differences in the recording device...). In this work we focus on recording device variability. Most studies refer to a single recording device for all of their voice samples, while those showing a variety of devices, it is due to the use of a variety of databases. As a consequence, they do not offer isolation of a single element on the channel, since recording environments are markedly different. This leads to lack of generalization, a common problem in machine learning known as domain adaptation. Training datasets are often small compared to the target population, and testing data sources do not often match training data [18].

These differences can even cause an unnoticed bias, leading to unwanted discrimination [19]. However, little effort has been made in studying the variability induced by differences in the communication channel. To the authors' best knowledge, [20] is the only published study about the robustness of telemonitoring systems against the impact of such differences, in this case mobile telephony network. In fact, it points out the need of a detailed microphone comparison.

In the present study we isolate the recording device and focus the attention on the effects of this element of the communication channel. The research hypothesis is that the recording setup has significant impact on the outcome of

the classifier, especially if the training process is made using a dataset obtained with a different setup than that used to record new unseen samples.

First, we study robustness against recording device variability of an automatic detection aid system. Then, we propose multicondition training (MCT) [21] as a useful generalization technique: we test its ability to improve robustness against differences between training and application devices. We show that this technique increases the ability of the machine learning model to distinguish healthy from PD diagnosed subjects when tested against previously unseen recording devices.

2. Materials and methods

The influence of smartphone as recording device has been studied by means of simulation. We used an in-house collected voice database, and designed a methodology to add the recording behavior of an assortment of smartphones to the recordings. Later, we trained a machine learning classifier to test the differences on classification accuracy due to the smartphone change. The following subsections provide details on each element of the experiment.

2.1. Participants

60 subjects volunteered for the experiment: 30 of them affected with PD, and 30 of them healthy. This number is in line with other research on PD assessment using vocal recordings [22, 23]. All subjects affected by PD were recruited in collaboration with the *Asociación Regional de Parkinson de Extremadura (ARPE)*. The inclusion criteria were that all of them should have been formally diagnosed with PD, and that their medical reports were available.

Healthy subjects were later recruited to approximately match the age and sex distribution of PD patients with the only requirement of neither having been diagnosed with PD nor having PD related symptomatology at recording time. All participants signed an informed consent. The research protocol was supervised and approved by the Bioethics Committee of the *Universidad de Extremadura*.

The group of people with PD is composed of 24 men and 6 women, with mean (standard deviation) age of 70.27 (9.54). The time since diagnosis was 9.93 (6.16) years. The Hoehn and Yahr stages ranged from 2 to 3, with a median stage of 2.5, i.e., patients in a mild-to-moderate condition. All the subjects were medicated with levodopa and the mean time since the last intake was 2.21 (1.32) hours.

2.2. Vocal task and recording equipment

For each subject included in the study, three different recordings were performed in a single session. Subjects were asked to vocalize an open \a\, for at least 5 seconds, as steadily as possible in both pitch and volume. Open \a\ is commonly used in research on automatic detection and assessment systems for voice impairment related diseases. Its ubiquity throughout different languages, and the simplicity of the experimental settings involved are the main reasons [10, 24].

The voice recordings were made using the same smartphone, model BQ Aquaris V, at a same sampling frequency of 44.1 kHz, and resolution of 16

bits. The setup for each recording session was the same: the distance from the speaker's mouth was about 30 cm, with the smartphone horizontally held, touchscreen up, and oriented so that the microphone points directly towards the source.

All the recordings were performed in the same room at the ARPE facilities under similar acoustical conditions. The room was not acoustically treated, although at recording time it was quiet. A trained person was present at all times in each recording session to ensure that all the participants properly followed the protocol, and to register the required complementary information.

Before any computation was made, all of the recordings were trimmed down in order to eliminate any leading or trailing silence. Also, one second segments were used to extract the considered speech features (see section 2.4), a duration deemed long enough [25]. This process was performed using Audacity software (release 2.0.5).

2.3. Recording device simulation

Different devices record the same sound in a disparate way, given the dissimilarities in design, component selection, and construction. The divergence can go from subtle, when comparing two specimens of the same model, to wide, when comparing models from different manufacturers, age, price range, or other features.

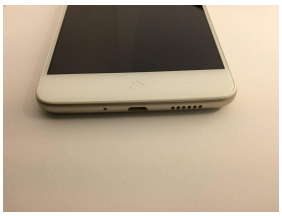
For our purposes, the ideal situation would be being able to record the same vocalization simultaneously using as many smartphones as possible. However, this situation is far from feasible: two devices can not be in the same position, and location influences the voice acquisition process since the human voice is not omnidirectional [26]. Furthermore, to the authors' best knowledge, database collection for any PD study has not considered the recording device variability problem.

Smartphone influence goes far beyond pure microphone frequency response. The recording system of a modern smartphone might include signal processing such as noise cancellation, compression or equalization. However, vendors do not offer information on their recording stacks. For that reason, recording device simulation seems to be an adequate alternative.

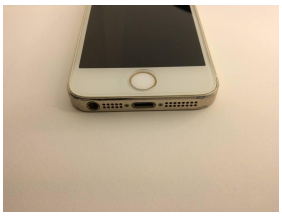
Having access to the original recording device, and to an assortment of smartphones, we can experimentally determine their individual frequency responses. We can process the recordings so that we subtract the effects of the original device, and estimate what the recording would have been if some another device had been used instead.

For the smartphone simulation, eight different devices were considered: Apple iPhone (model A1533); Apple iPhone S and Apple iPhone S(2) (model A1688), without and with an external battery attached respectively, which alters the microphone opening; iPhone SE (model A2296); OnePlus Nord (model AC2003); Realme 8 (model RMX3241); Redmi Note 9 Pro (model M2003J6B2G); and Samsung A51 (model SM-A515F/DSN). The selection criteria was having a variety of manufacturers, with high market penetration and relatively affordable. Microphone placement for each of them is shown in Fig. 1.

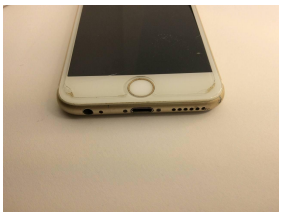
We followed the IEC 60268-4:2018 standard for microphone testing to the extent possible. It describes the way a microphone should be tested in order to obtain its characteristics, including frequency response and directional pattern.



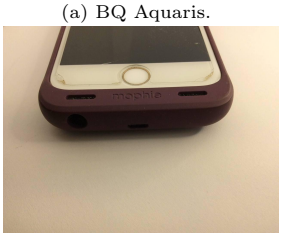
(a) BQ Aquaris.



(b) Apple iPhone.



(c) Apple iPhone S.



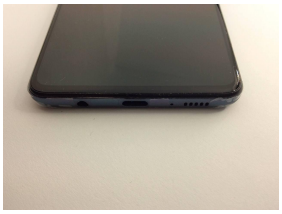
(d) Apple iPhone S(2).

(e) Apple iPhone SE.

(f) OnePlus Nord.

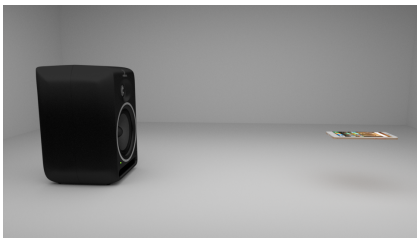
(g) Realme 8.

(h) Redmi Note 9.

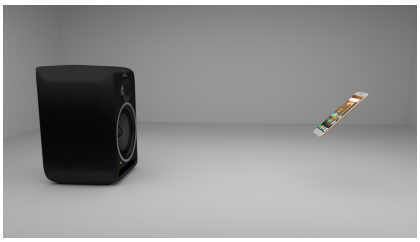


(i) Samsung A51.

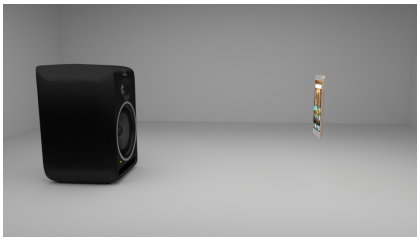
Figure 1: Microphone placement for all the devices.



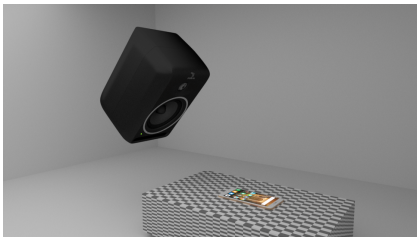
(a) Position 1. Horizontally suspended.



(b) Position 2. 45° slant suspended.



(c) Position 3. Vertically suspended.



(d) Position 4. On a table.

Figure 2: Test setup for each source-smartphone positioning.

The standard is intended for stand alone microphones, thus not all requirements could be fulfilled since smartphone recording is a black box where processing is unknown.

Testing was made in an anechoic chamber located at Array Processing Lab

(*Universidad de Valladolid*). The loudspeaker model was Hedd Audio Type 07, which has a frequency response of -3 dB in the range 38-40,000 Hz. As for the reference microphone, we used Behringer ECM8000, with a frequency response of -3 dB in the range 20-20,000 Hz. The frequency swipes and recordings were made using Audacity software (release 2.0.5) and the sound interface was TASCAM US-322.

A total of 4 different orientations were tested: 3 different pitch angles (rotation around X-axis): 0°, position 1, Fig. 2a; 45°, position 2, Fig. 2b; and 90°, position 3, Fig. 2c. The device microphone as the center point for rotation was always considered, thus microphone placement relative to the sound source remained the same. Also, an extra position was tested by placing the smartphone on a horizontal surface as could be a table, which is not considered in the standard, Fig. 2d.

For positions 1-3 the sound source was located at a distance of 30 cm from the microphone and, due to the technical difficulties of placing the reference microphone and the device under test in the exact same spot in space, substitution method was discarded and simultaneous comparison method was used. For position 4, distance to the sound source was located 20 cm over the horizontal plane where the smartphone is lying, and the distance to the source was still 30 cm. In this case, substitution method was used, placing the reference microphone without the horizontal surface present in the same spot as the smartphone microphone would later be placed.

A continuous frequency sweep was performed in the 0-22,050 Hz range which, along with distance to the source, follows IEC 60268-4:2018 standard. Fig. 3 shows the magnitude of the Fourier Transform for a sample recording using each device simulation in each position.

Given that 1 second length recordings were used, at a sampling rate of 44,100 Hz, frequency responses were obtained with a resolution of 1 Hz. The frequency gains for BQ Aquaris-V were subtracted from each recording spectral analysis to obtain a "clean" recording without device influence. Then, frequency gains for each device were added so we could simulate each device influence. The gains were applied by transforming the signal to frequency domain using Fourier Transform, operating with the gains obtained for each device, and reconstructing the signal by means of Inverse Fourier Transform.

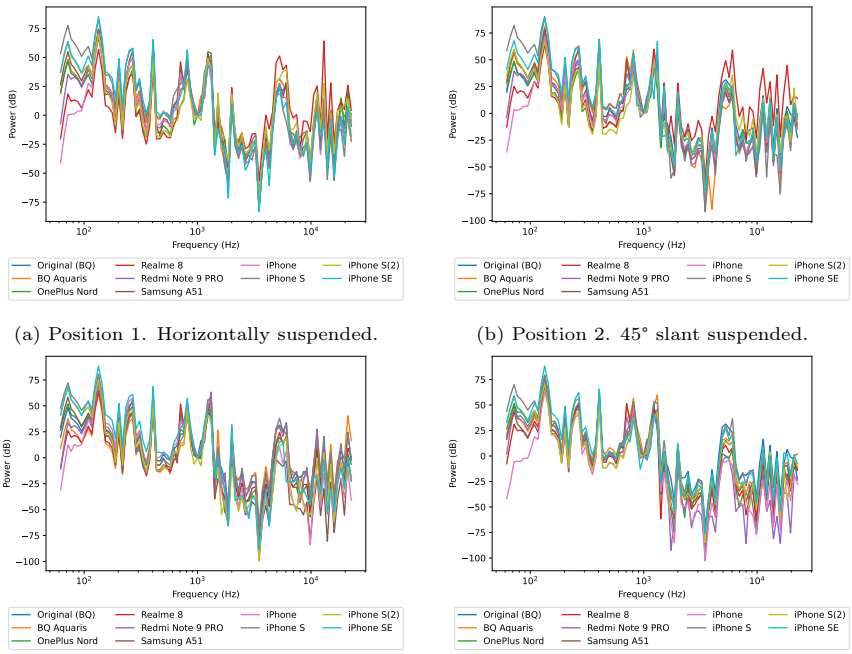
2.4. Feature extraction

35 features were initially considered, including Cepstral Peak Prominence, Correlation Dimension, First Minimum in Mutual Information, Glottal to Noise Excitation, Harmonic to Noise Ratio, Hurst's Exponent, Jitter, Lempel Ziv Complexity, Mel Frequency Cepstral Coefficients, Multifractal Spectrum Width, Permutation Entropy, Pitch Period Entropy, Recurrence Period Density Entropy, Shannon's Entropy, Shimmer, and Zero Crossing Rate.

More detailed information on the considered features can be found in [12, 27]. They have been widely used in studies on pathological speech since they measure different speech impairment aspects. Also, a feature selection process is performed (see Subsection 2.5) to select and employ only the most useful ones.

2.5. Variable selection and classification

As stated in section 2.2, three vocal samples were collected from each participant. Those samples were individually processed, simulating 8 additional



(a) Position 1. Horizontally suspended.

(b) Position 2. 45° slant suspended.

(c) Position 3. Vertically suspended.

(d) Position 4. On a table.

Figure 3: Test setup for each source-smartphone positioning.

devices in 4 positions, and totaling 36 device/position combinations. Features were extracted for all the subjects, obtaining a data matrix of 180 voice samples $\times$ 35 acoustical features for each combination. Later, the values for each feature were averaged per patient, reducing the matrix size to 60 $\times$ 35 and the result was used as input data.

Different experiments were conducted by changing train and test datasets. For each device/position combination we built a machine learning model consisting in three steps: feature selection, grid search, and classification. The goal was to maximize accuracy. This process follows the methodology shown in [27]. In this case, we used a passive aggressive classifier because it yielded high accuracy levels and low computation times in early research stages.

The initial number of features is large compared to the number of voice samples, with a ratio close to 1/2. This could lead to overfitting problems in the train phase, and limited statistics for valuable results. Therefore, reducing the number of features is a critical step in the pipeline.

Feature selection was performed using Recursive Feature Elimination with Cross Validation (RFECV). This technique obtains an optimal size smaller feature subset by discarding the least important features in an iterative process. This was repeated 500 times using a 2-fold cross validation scheme. The number of repetitions was chosen looking for stability, and is based in the law of large numbers (LNN), which states that the larger the number of trials, the closer to the expected value the average will be. The resulting subsets were stored and used in a sorting mechanism so that the features could be ranked by their prevalence in the selections. This ranking would be later used to obtain a small

working subset.

Grid search looks for the best hyperparameters given a dataset and a classifier. A passive aggressive classifier was used, matching the one used in the feature selection step, and training data was set to match feature selection as well. At this point, the system must be oblivious to the testing samples to be used.

Finally, for each train/test device and position combination, we found the feature subset that yielded the best accuracy for that specific combination, which brought to an end the feature selection process. Those subsets are small compared to the initial one, therefore small compared to the dataset size, alleviating the feature number to sample size ratio problem. To obtain these small sets we trained the classifier 35 times using the n most important features according the rank obtained in the feature selection phase, being $n = 1 \dots 35$. Finally, we found which classifier yielded the highest accuracy, thus finding a small feature subset for each cross validation split. This was repeated for each train/test device and position combination.

We used a stratified shuffle and split strategy as model validation and generalization technique. Cross validation enables us to generalize the performance of the machine learning model and obtain an estimate of the classification accuracy if tested against new samples, as long as those samples come from a dataset with similar statistical characteristics. We are looking specifically towards the influence of recording devices on the outcomes. Therefore, it is of interest to maintain every other constraint constant. Stratification ensures that the proportion of healthy/pathological subjects is constant across training and testing sets. The number of splits was 1000, with a 2/3 (40 subjects) training size and 1/3 (20 subjects) testing size. The number of splits was chosen based on LNN.

Also, random sampling was conditioned so the *n-th* split for each phone-position combination selects the same individuals for testing and training, thus the differences in accuracy can only be attributed to splitting.

2.6. Multicondition training

Whenever a classifier is built with samples from a single data acquisition setup, the system may specialize in the environmental characteristics of the experimental setup, and may lack accuracy if tested with samples from other sources. Therefore, it is necessary to develop strategies to avoid this problem. MCT, which takes into consideration the variability of acquisition conditions in the training dataset [21], has proven to be useful to improve robustness of classifiers for Reinke's space diseases in an environment affected by noise [28].

Differences in the recording equipment and its relative position to the subject may also degrade the performance when this source of variability has not been taken into account in the training process. The robustness of a multicondition trained classifier based on the different smartphone frequency responses has been tested. The performance of this system is compared with the aggregated results obtained when training with a smartphone recordings and testing with a different smartphone; the aimed improvement should be assessed not only in the mean accuracy, but also in the dispersion of the results obtained.

Among the different multicondition strategies available, [28] has shown that asymmetry (using only one recording per subject independently of the number of recording conditions present in the dataset) is the right strategy.

Training phase for MCT follows the same schema than single condition training: A feature selection phase, followed by grid search and classification. Two different approaches were studied: First, each recording in the train set was affected by a randomly selected device, and all the recordings in the test set were affected by the same device, named all to one; secondly, both train and test sets were affected by a randomly selected device, named all to all. The devices were chosen so that the proportion of recordings affected by each device was constant, with the limitation of split size and number of devices being coprime integers.

Finally, the *umpteenth* split for each cross validation step selects the same individuals for each set, which are also the same individuals for the *n-th* split in single condition training (see Subsection 2.5). Thus, the differences in results between steps in cross validation are to be attributed only to splitting, and the differences between single, all to one MCT, and all to all MCT, can be attributed merely to the training strategy.

2.7. Statistical analysis

Descriptive statistics such as mean, standard deviation, and coefficient of variation have been considered. Coefficient of variation is a dimensionless relative dispersion measure that is defined as $CV = s/\bar{x}$, where s stands for standard deviation and $\bar{x}$ for mean. Statistical hypothesis tests have been used to report statistically significant differences between groups. When normality condition could be assumed, unpaired t-tests for the homoskedastic and heteroskedastic cases were applied because of their statistical power [29]. Otherwise, the non-parametric counterpart (Mann-Whitney U test) was applied [30]. Both tests provide a p-value that can be thought of as the probability of finding the data under the assumption of the null hypothesis, i.e. under the hypothesis of no difference between groups. P-values lower than 0.05 reported statistically significant differences.

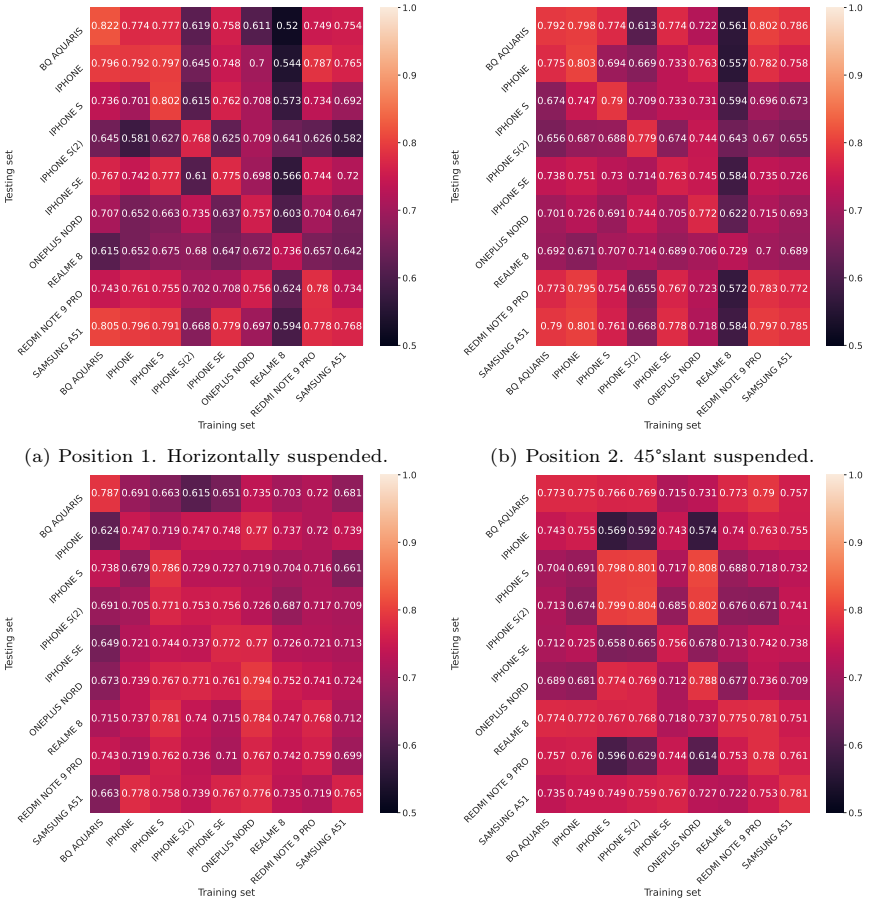
3. Results and discussion

We have studied the influence that changing the recording device and its relative position to the subject might have on the performance of the system.

Fig. 4 shows the accuracy obtained for all the combinations for training device, testing device, and position, as discussed in section 2.3. It is noticeable how the BQ Aquaris-BQ Aquaris-Position 1 combination yields the best accuracy overall (0.822). The original recording device and experimental setup should be expected to get the best results since the recordings have not been processed and nothing had to be simulated.

Not every position shows an even behavior. Fig. 4c shows a higher homogeneity in results for position 3, with a more equal "heat" across all training-testing combinations than Figs. 4a, 4b, and 4d. For a smartphone common use case, the microphone points towards speaker's face when the user is making a phone call; that direction is normal to the smartphone screen so it is pointing towards sound source in position 3 experiments, which explains the good results.

Position 4 yields surprising results, overperforming positions 1-3 in 49, 43, and 47 out of 81 combinations, respectively. It is commonly advised for any measurement procedure where microphones are involved that the microphone



(a) Position 1. Horizontally suspended.

(b) Position 2. 45°slant suspended.

(c) Position 3. Vertically suspended.

(d) Position 4. On a table.

Figure 4: Classification accuracy obtained with every training-testing device combination for each of the device positions considered.

should be far enough from reflective surfaces [31]. However, in our experiments, placing the phone on a horizontal surface shows better behavior than positions 1 and 2.

Table 1 is consistent with this analysis. It shows the mean value and coefficient of variation for all the accuracies shown in each Fig. 4 subfigure. Mean accuracy increases from position 1 to position 2, and from position 2 to position 3, while the coefficient of variation decreases. This explains the homogeneity perceived for position 3 where the "heat" seems more evenly distributed in the subfigure. Also, accuracy values are higher in general.

	Position			
	1	2	3	4
Average	0.701	0.718	0.729	0.731
CV	0.103	0.085	0.052	0.073

Table 1: Average accuracy and coefficient of variation for each subfigure in Fig. 4, training and testing with a single device.

Position	Type	N	Mean	Stand. Dev.	P-value
Position 1	Matched	9	0.778	0.025	<0.001
	Mismatched	72	0.692	0.071	
Position 2	Matched	9	0.777	0.022	<0.001
	Mismatched	72	0.710	0.061	
Position 3	Matched	9	0.768	0.018	<0.001
	Mismatched	72	0.724	0.037	
Position 4	Matched	9	0.779	0.017	<0.001
	Mismatched	72	0.725	0.053	

Table 2: Count, mean, and standard deviation for one-one comparison.

Position 4 shows higher mean accuracy than position 3, and positions 1 and 2 consequently. However, the coefficient of variation is higher than that of position 3, showing a slight advantage for the latter, while it is still lower than the coefficients of positions 1 and 2. This places position 4 as the second best setup, very close to position 3. Given the sound source position relative to the microphone, the smartphone angle is $\alpha = \arcsin(20/30) \approx 42°$, close to that of position 2. The reason for this improvement is not clear: reflections on the surface and resonances should be accounted for, and, instinctively, one might expect a degradation in performance, but the data shows the opposite.

Every combination other than BQ Aquaris-BQ Aquaris should be affected by the simulation, with some undetermined side effects than might induce error into the system. However, looking at the main diagonal in Figs.4a-4d, where the training and testing recording/simulated device is the same, it is shown that the performance of all the systems is similar regardless its position, showing accuracies in the 0.73-0.82 range. This combination is always at least a 89% of the BQ Aquaris-BQ Aquaris-Position 1 combination, thus retaining most of the classifying ability.

Furthermore, the matched diagonal elements seem to yield better results than mismatched experiments. This is supported by the statistical analysis shown in Fig. 5 and Table 2. Mean accuracy for matched devices is almost even across all positions, in the vicinity of 0.77, whereas mismatched experiments lose between 11% and 7% depending on the position. Error bars shown in Fig. 5 underline the improving effect of matched over mismatched experiments. A hypothesis test has also been applied. The results reveal that in all positions there exist statistically significant differences in accuracies between matched and mismatched conditions, being the values for mismatched lower. All p-values were lower than 0.001.

Based on the matched-mismatched differences in accuracy, we proceeded to train the system under an MCT schema, testing its capacity to improve the system robustness. Fig. 6b shows results for the experiments carried out: For each position we train the classifier with a mixture of multiple recording devices and test their abilities against an individual device recordings. It is worth noting that for MCT the train/test split is stratified in both PD/healthy ratio and recording device prevalence, so differences in results can be attributed, like in single condition experiments, to the patients selected for each cross validation split.

For comparison purposes, we show in Fig. 6a the row-wise mean accuracy

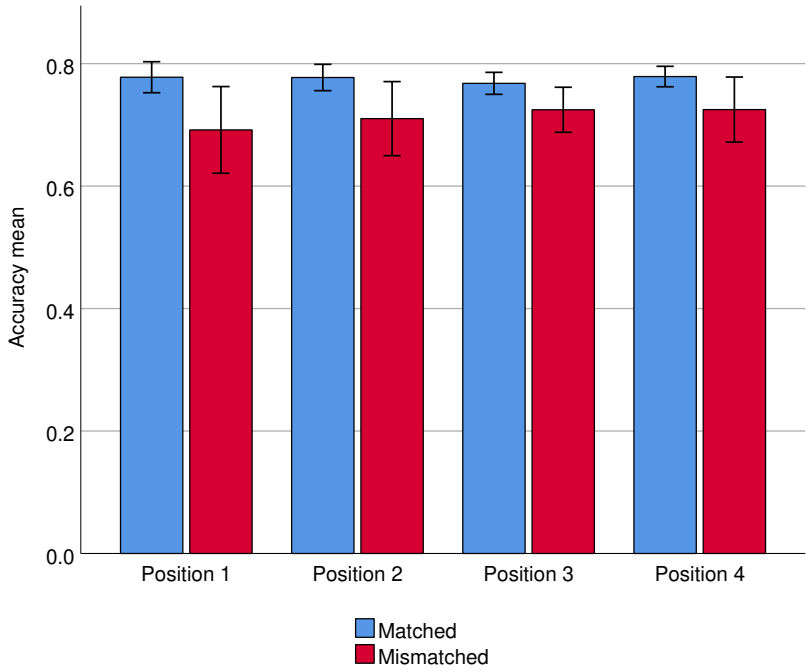


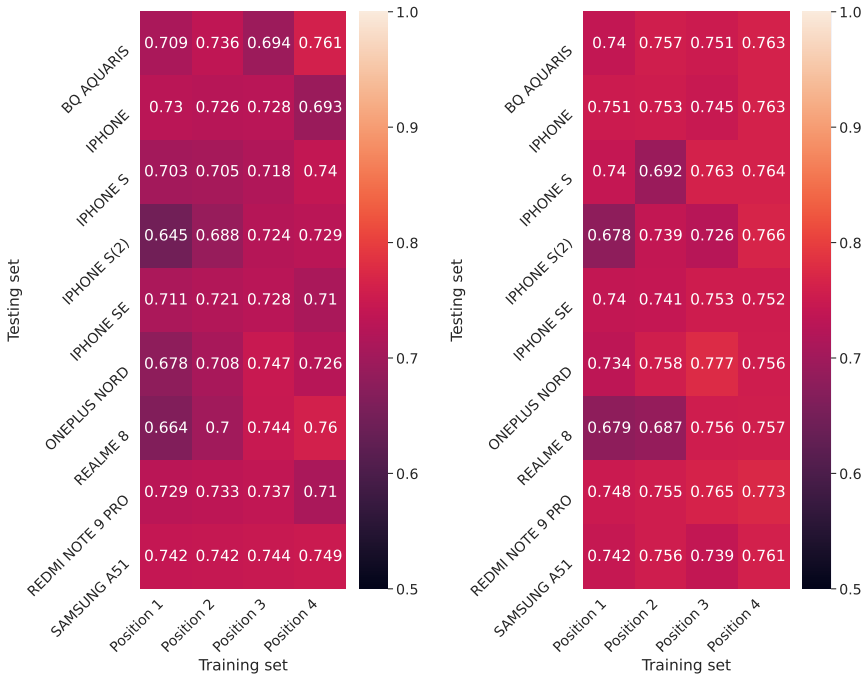
Figure 5: Classification accuracy obtained with every training-testing device combination for each of the device positions considered.

obtained in Figs.4a-4d. Each column summarizes results for single phones as test devices in a position. We can see that in this situation MCT improves the mean classifier performance for each position and for each recording device under test, since the results exceed the equivalent mean values for all the combinations. Exceptions are found in position 2-iPhone S(2), position 2-Realme 8, position 3-Samsung A51, and position 4-Realme 8 simulations. However, these exceptions in MCT barely underperform a 2% from the mean, whilst the mean improvement in accuracy due to MCT is about 4%, with peaks of 8%.

The fact that MCT gets better average results and lower error (Fig. 7) points out that MCT might contribute to build more robust systems. This is also backed by the statistical analysis results shown in Table 3: the difference between MCT and SC is more than one standard deviation apart and shows statistically significant differences (p-value < 0.05). This underscores the MCT usefulness to improve robustness. Position 2 is an exception to this, although its p-value is 0.058, very close to statistical significance.

Comparing position performance, both position 1 and position 2 seem more homogeneous with MCT (Fig. 6b) than they do without using it (Fig. 4a). Also, the growing trend of accuracies for positions 1, 2, and 3 appears to remain with MCT, and position 4 still rivals with position 3 results. This qualitative analysis is supported by Table 4: Compared with Table 1, averages obtained per position are higher with MCT in all cases, and coefficients of variation are lower as well. It is remarkable how, in this case, position 4 yields the best results, beating those obtained for position 3.

Finally, Table 5 shows the results obtained for an all to all MCT experiment (using a train set and a test set built with a mixture of all recording devices).



(a) Row-wise mean (mean per testing device) of the accuracies of each subfigure in Fig. 4.

(b) Mean accuracy for MCT in each position.

Figure 6: Comparison between mean accuracy obtained attending to position. Mean testing device accuracy versus MCT mean accuracy.

Position	Type	N	Mean	Stand. Dev.	P-value
Position 1	SC	9	0.701	0.033	0.040
	MCT	9	0.728	0.028	
Position 2	SC	9	0.718	0.018	0.058
	MCT	9	0.738	0.028	
Position 3	SC	9	0.729	0.017	0.006
	MCT	9	0.753	0.015	
Position 4	SC	9	0.731	0.024	0.004
	MCT	9	0.762	0.006	

Table 3: Count, mean, standard deviation, and p-value for MCT and SC.

	Position			
	1	2	3	4
Average	0.728	0.738	0.753	0.762
CV	0.039	0.038	0.020	0.008

Table 4: Average accuracy and coefficient of variation for each column in Fig. 6b, MCT with all the devices and testing with a single device.

The average accuracy values for positions 1-4 are consistent to the mean accuracy shown in Table 4 as should be expected: the low all to one CV values suggest that an all to all experiment should yield an average close to the mean

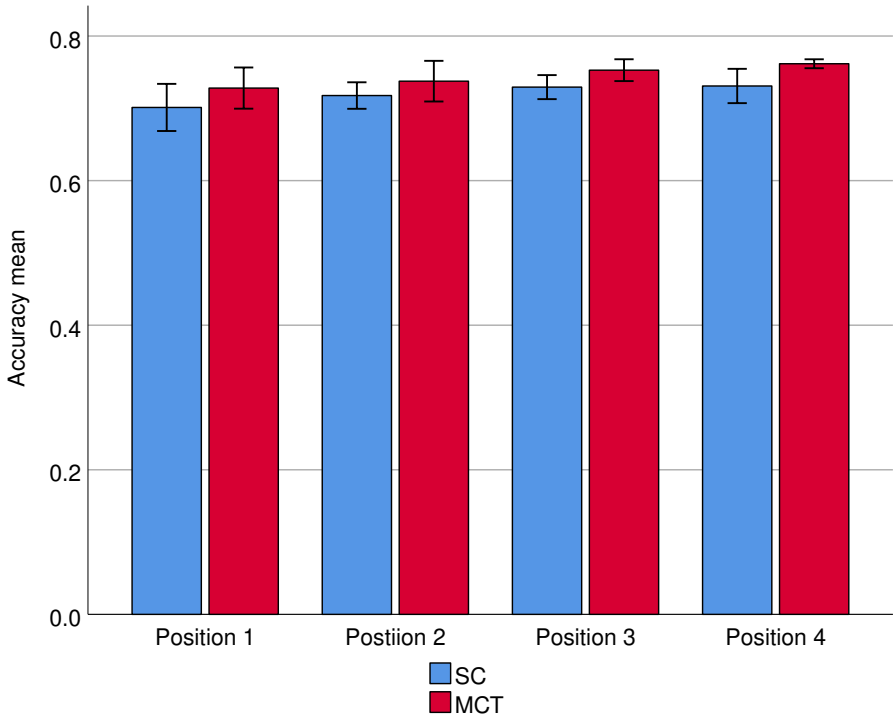


Figure 7: Classification accuracy obtained with every training-testing device combination for each of the device positions considered.

	Position			
	1	2	3	4
Average	0.724	0.734	0.754	0.762

Table 5: Mean accuracies for MCT, all devices in training and testing sets.

average of the all to one experiments, which is the case. In fact, the difference of Table 5 values from those shown in Table 4 is lower than 0.01% in all cases, and specifically lower than 0.001% for positions 3 and 4. Those results lie within error in positions 3 and 4, which are consistently yielding the most stable results.

For the sake of completeness, since in diagnostic tools sensitivity and specificity are important metrics, we also include their values as supplementary data. In the case of MCT, the mean sensitivity (specificity) values obtained after testing will all the devices, are 0.737 (0.711), 0.744 (0.724), 0.772 (0.737), 0.782 (0.743), for positions 1, 2 ,3 and 4, respectively. Therefore, the proposed MCT system is more sensitive than specific. In the case of a screening test for a disease, the role of sensitivity is more critical than that of specificity

Although the goal of this study is to analyze the effects of changing the recording device, accuracy obtained for the realistic scenario, where we use BQ Aquaris smartphone for both train and test phases, is on par with related literature. For example, [32] uses a variety of phonemes from PC-Gita and Viswanthan's databases for PD detection. Using \a\ phoneme it reaches an accuracy

of 0.693 and 0.858, respectively. On its side, [33] uses Neurovoz, ItalianPVS and mPower databases with accuracies of 0.854, 0.990, and 0.754, respectively, always using \a\ phoneme.

There have been some efforts in studying the reliability of smartphones as a source of data for voice health assessment [34, 35]. They show that certain features are more affected than others, and consider smartphones as a valid recording device. They compare the error of an assorted set of smartphones performance against a studio microphone, but do not get into the automatic assessment phases.

Regarding selected features, there exists a high variability on experimental conditions: 9 single condition feature selection experiments per position; 1 multicondition feature selection experiment per position in all to one configuration; and 1 in all to all configuration. However, an examination of the most selected features shows that, for each experiment, the 10 most used features are a subset of the following: Glottal to Noise Excitation, Lempel Ziv Complexity, Mel Frequency Cepstral Coefficients 3, 5, 6, 9, 10, 11, 13, Cepstral Peak Prominence, and First Zero in Correlation Function.

Those features are usually considered in scientific literature [12]. Some of them also stand out as reliable: in [33], experiments with different databases show that MFCCs are usually ranked among the most important features they considered. The prevalence of the aforementioned features across every experiment shows that, for PD, these might be the most robust ones among the features considered, which should be further investigated.

All of the experiments were designed having in mind that future health telemonitoring, specifically voice assessment, and particularly PD diagnosis and monitoring, will probably be linked to the development of smartphones and their capabilities. Many efforts have been already made, like mPower initiative [36], recruiting volunteers and recording their voices among other motor and cognitive tests for PD monitoring, or Parkinson Voice Initiative [37], collecting telephone-quality recordings from subjects from seven different countries. In the case of PD, [20] suggests the necessity of a detailed comparison of different microphones from different smartphones to complete the analysis of the full communication path in a hypothetical telemonitoring system.

This paper fills that gap. Results obtained in matched conditions show that most modern smartphones provide adequate recording systems for this particular application. Furthermore, the quality of the recording device is not nearly as important as a right setup for experimentation. It is worth stressing that in this paper we do not intend to recommend a specific recording device, but to underline the importance of training with a variety of sound sources. Environmental influence has been tested in previous work [27, 28]. The present paper complements them in channel description even though those studies revolve around voice conditions other than PD.

However, the results can be transposed to any other condition. The experiments test the influence of recording device and their positioning in the outcome of a statistical learning algorithm. The fact that we can compare positioning of the same device allows us to discard any other influential factor, since the experimental setup fully isolates the considered variables. The differences in simulation between two different positions given a smartphone, or between two smartphones having selected a position, can only be attributed to that change, as the anechoic chamber eliminates any noise source other than those inherent

to the recording system, and the ones already present in the original recording.

Moreover, the present results can be further extended to any other telemonitoring setup. In this paper, we have considered .wav lossless voice recordings. Other efforts in telemedicine development use real time connections. To the authors' best knowledge, influence of other channels than that of cell phone networks have not been tested. However, there is a wide range of commercial voice over IP solutions, and it is a hot topic in communications development mostly due to current teleworking needs. All of these solutions will necessarily be placed after voice sampling, and therefore the recording setup would have an influence on them all, whether it is live or recorded.

4. Conclusions

We have studied the effects of smartphone selection and placement in the accuracy of an automatic detection aid system for PD based on voice recordings. Experimental results indicate that it is a good practice to test the system using recordings obtained with the same device used for testing. If we acknowledge the variability in recording devices used for a widespread technology, results may vary. Differences up to 37% were found when using other smartphone than that used for training.

We have also proposed a methodology to overcome the limitation in recording device selection by using MCT. This technique offers lower results dispersion with an increase in accuracy compared to the averaged results of single condition. However, further studies would be required to increase the statistical power of the results, involving a higher number of voice samples.

Also, we have found that recording device position relative to the speaker has a high impact on results. Holding the phone vertically right in front of the speaker yields the best results, and placing the phone atop a table is the second best option.

Acknowledgments

This research is part of R&D&I Projects PID2021-122209OB-C32 and MTM2017-86875-C3-2-R, funded by MCIN/AEI/10.13039/501100011033/; Grants GR21057 and GR21072, funded by Junta de Extremadura and the European Regional Development Fund (ERDF/FEDER); and Grant FPU18/03274 (Ministerio de Universidades).

References

[1] O.-B. Tysnes, A. Storstein, Epidemiology of Parkinson's disease, Journal of neural transmission 124 (8) (2017) 901–905. doi:10.1007/s00702-017-1686-y.

[2] Z. Ou, J. Pan, S. Tang, D. Duan, D. Yu, H. Nong, Z. Wang, Global trends in the incidence, prevalence, and years lived with disability of Parkinson's disease in 204 countries/territories from 1990 to 2019, Frontiers in public health (2021) 1994.

[3] S. L. Oh, Y. Hagiwara, U. Raghavendra, R. Yuvaraj, N. Arunkumar, M. Murugappan, U. R. Acharya, A deep learning approach for Parkinson's disease diagnosis from EEG signals, Neural Computing and Applications 32 (15) (2020) 10927–10933. doi:10.1007/s00521-018-3689-5.

[4] N. Amoroso, M. La Rocca, A. Monaco, R. Bellotti, S. Tangaro, Complex networks reveal early MRI markers of Parkinson's disease, Medical image analysis 48 (2018) 12–24. doi:10.1016/j.media.2018.05.004.

[5] M. Belić, V. Bobić, M. Badža, N. Šolaja, M. Đurić-Jovičić, V. S. Kostić, Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease—a review, Clinical neurology and neurosurgery 184 (2019) 105442. doi:10.1016/j.clineuro.2019.105442.

[6] W. Pawlukowska, A. Szylińska, D. Kotlęga, I. Rotter, P. Nowacki, Differences between subjective and objective assessment of speech deficiency in Parkinson disease, Journal of Voice 32 (6) (2018) 715–722. doi:https://doi.org/10.1016/j.jvoice.2017.08.018.

[7] T. Zhang, Y. Zhang, H. Sun, H. Shan, Parkinson disease detection using energy direction features based on EMD from voice signal, Biocybernetics and Biomedical Engineering 41 (1) (2021) 127–141. doi:10.1016/j.bbe.2020.12.009.

[8] W. Rahman, S. Lee, M. S. Islam, V. N. Antony, H. Ratnu, M. R. Ali, A. A. Mamun, E. Wagner, S. Jensen-Roberts, E. Waddell, T. Myers, M. Pawlik, J. Soto, M. Coffey, A. Sarkar, R. Schneider, C. Tarolli, K. Lizarraga, J. Adams, M. A. Little, E. R. Dorsey, E. Hoque, Detecting Parkinson disease using a web-based speech task: Observational study, Journal of Medical Internet Research 23 (10) (2021) e26305. doi:10.2196/26305.

[9] D. Montaña, Y. Campos-Roca, C. J. Pérez, A diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson's disease, Computer Methods and Programs in Biomedicine 154 (2018) 89–97. doi:https://doi.org/10.1016/j.cmpb.2017.11.010.

[10] G. Solana-Lavalle, J.-C. Galán-Hernández, R. Rosas-Romero, Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features, Biocybernetics and Biomedical Engineering 40 (1) (2020) 505–516. doi:10.1016/j.bbe.2020.01.003.

[11] M. Hireš, M. Gazda, P. Drotár, N. D. Pah, M. A. Motin, D. K. Kumar, Convolutional neural network ensemble for Parkinson's disease detection from voice recordings, Computers in Biology and Medicine (2021) 105021doi:10.1016/j.compbiomed.2021.105021.

[12] J. A. Gómez-García, L. Moro-Velázquez, J. I. Godino-Llorente, On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art, Biomedical Signal Processing and Control 51 (2019) 181–199. doi:10.1016/j.bspc.2018.12.024.

[13] J. A. Gómez-García, L. Moro-Velázquez, J. I. Godino-Llorente, On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors, Biomedical Signal Processing and Control 48 (2019) 128–143. doi:10.1016/j.bspc.2018.09.003.

[14] A. Tsanas, Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning, Ph.D. thesis, University of Oxford UK, D. Phil. Thesis (2012).

[15] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, P. Klumpp, P. A. Pérez-Toro, D. Escobar-Grisales, N. Roth, C. D. Ríos-Urrego, M. Strauss, H. A. Carvajal-Castaño, S. Bayerl, L. R. Castrillón-Osorio, T. Arias-Vergara, A. Künderle, F. O. López-Pabón, L. F. Parra-Gallego, B. Eskofier, L. F. Gómez-Gómez, M. Schuster, E. Nöth, Apkinson: the smartphone application for telemonitoring Parkinson's patients through speech, gait and hands movement, Neurodegenerative Disease Management 10 (3) (2020) 137–157. doi:10.2217/nmt-2019-0037.

[16] H. Yoon, N. Gaw, A novel multi-task linear mixed model for smartphone-based telemonitoring, Expert Systems with Applications 164 (2021) 113809. doi:10.1016/j.eswa.2020.113809.

[17] J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, Robust automatic speech recognition: a bridge to practical applications, Academic Press (2015).

[18] W. M. Kouw, M. Loog, A review of domain adaptation without target labels, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (3) (2021) 766–785. doi:10.1109/TPAMI.2019.2945942.

[19] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, M. H. Chin, Ensuring fairness in machine learning to advance health equity, Annals of internal medicine 169 (12) (2018) 866–872. doi:10.7326/M18-1990.

[20] A. Tsanas, M. A. Little, L. O. Ramig, Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network, IEEE Access 9 (2021) 11024–11036. doi:10.1109/ACCESS.2021.3050524.

[21] D. Garcia-Romero, X. Zhou, C. Y. Espy-Wilson, Multicondition training of gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 4257–4260. doi:10.1109/ICASSP.2012.6288859.

[22] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, N. Dehak, A forced gaussians based methodology for the differential evaluation of Parkinson's disease by means of speech processing, Biomedical Signal Processing and Control 48 (2019) 205–220.

[23] M. Novotný, P. Dušek, I. Daly, E. Růžička, J. Rusz, Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with Parkinson's

disease: correlation between acoustic speech characteristics and non-speech motor performance, Biomedical Signal Processing and Control 57 (2020) 101818.

[24] L. Naranjo, C. J. Perez, Y. Campos-Roca, J. Martin, Addressing voice recording replications for Parkinson's disease detection, Expert Systems with Applications 46 (2016) 286–292.

[25] A. J. Romann, B. C. Beber, C. A. Cielo, C. R. d. M. Rieder, Acoustic voice modifications in individuals with Parkinson disease submitted to deep brain stimulation, International archives of otorhinolaryngology 23 (2019) 203–208. doi:10.1055/s-0038-1675392.

[26] C. Pörschmann, J. M. Arend, Investigating phoneme-dependencies of spherical voice directivity patterns, The Journal of the Acoustical Society of America 149 (6) (2021) 4553–4564. doi:10.1121/10.0005401.

[27] M. Madruga, Y. Campos-Roca, C. J. Pérez, Impact of noise on the performance of automatic systems for vocal fold lesions detection, Biocybernetics and Biomedical Engineering 41 (3) (2021) 1039–1056. doi:10.1016/j.bbe.2021.07.001.

[28] M. Madruga, Y. Campos-Roca, C. J. Perez, Multicondition training for noise-robust detection of benign vocal fold lesions from recorded speech, IEEE Access 9 (2020) 1707–1722. doi:10.1109/ACCESS.2020.3046873.

[29] G. D. Ruxton, The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test, Behavioral Ecology 17 (4) (2006) 688–690. doi:10.1093/beheco/ark016.

[30] P. E. McKnight, J. Najab, Mann-Whitney U Test, John Wiley & Sons, Ltd, 2010, pp. 1–1. doi:10.1002/9780470479216.corpsy0524.

[31] Acoustics - Measurement of room acoustic parameters - Part 1: Performance spaces, Standard, International Organization for Standardization, Geneva, CH (Jun. 2009).

[32] N. D. Pah, M. A. Motin, D. K. Kumar, Phonemes based detection of parkinson's disease for telehealth applications, Scientific Reports 12 (1) (2022) 1–9.

[33] A. S. Ozbolt, L. Moro-Velazquez, I. Lina, A. A. Butala, N. Dehak, Things to consider when automatically detecting Parkinson's disease using the phonation of sustained vowels: Analysis of methodological issues, Applied Sciences 12 (3) (2022) 991.

[34] F. Schaeffler, S. Jannetts, J. M. Beck, Reliability of clinical voice parameters captured with smartphones–measurements of added noise and spectral tilt, in: Proceedings of the 20th Annual Conference of the International Speech Communication Association INTERSPEECH, Graz, Austria, 15-19 September 2019, ISCA, 2019.

[35] S. Jannetts, F. Schaeffler, J. Beck, S. Cowen, Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types, International journal of language & communication disorders 54 (2) (2019) 292–305.

[36] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, A. D. Trister, The mPower study, Parkinson disease mobile data collected using researchkit, Scientific data 3 (1) (2016) 1–9. doi:10.1038/sdata.2016.11.

[37] S. Arora, L. Baghai-Ravary, A. Tsanas, Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice, The Journal of the Acoustical Society of America 145 (5) (2019) 2871–2884. doi:10.1121/1.5100272.