Evaluation of the neurological state of people with Parkinson's disease using i-vectors

N. Garcia[1]*, J. R. Orozco-Arroyave*[1,2]*, L. F. D'Haro*[3]*, Najim Dehak*[4]*, Elmar Nöth*[2]

[1]Faculty of Engineering, Universidad de Antioquia UdeA, Medellín, Colombia
[2]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[3]Institute for Infocomm Research I^2R, A*STAR, Singapore
[4]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA

nicanor.garcia@udea.edu.co

Abstract

The i-vector approach is used to model the speech of PD patients with the aim of assessing their condition. Features related to the articulation, phonation, and prosody dimensions of speech were used to train different i-vector extractors. Each i-vector extractor is trained using utterances from both PD patients and healthy controls. The i-vectors of the healthy control (HC) speakers are averaged to form a single i-vector that represents the HC group, i.e., the reference i-vector. A similar process is done to create a reference of the group with PD patients. Then the i-vectors of test speakers are compared to these reference i-vectors using the cosine distance. Three analyses are performed using this distance: classification between PD patients and HC, prediction of the neurological state of PD patients according to the MDS-UPDRS-III scale, and prediction of a modified version of the Frenchay Dysarthria Assessment. The Spearman's correlation between this cosine distance and the MDS-UPDRS-III scale was 0.63. These results show the suitability of this approach to monitor the neurological state of people with Parkinson's Disease.

Index Terms: Speech disorders, i-vectors, Parkinson's disease, cosine distance

1. Introduction

Parkinson's Disease (PD) is the second most common neurodegenerative disease worldwide after Alzheimer's and its incidence and prevalence are raising [1]. PD is often associated with its primary motor symptoms which include tremor, akinesia, bradykinesia, and postural instability [2]. These motor symptoms limit the mobility of patients and make it hard for them to attend medical appointments and therapy sessions [3]. Additionally, other motor symptoms include speech disorders that affect the oral communication skills of the patient [3, 2]. The majority of PD patients develop some sort of dysarthria during the course of the disease [4]. To follow the progression of the disease, neurologist experts apply different tests to the patient. The Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [5] is one of the most common. This perceptual scale is divided into four sections and assesses motor and non-motor abilities of PD patients. This study considers only the third section (MDS-UPDRS-III) which evaluates the motor capabilities of the patients in 33 items. However, speech is evaluated only in one of them. As the speech production is one of the most complex processes in the brain, it is worth to evaluate it with more detail. The Frenchay Dysarthria Assessment is a test administered to evaluate the level of dysarthria [6]. It evaluates the speech production process of the patient in eight sections considering the different systems involved in the speech production process, e.g., respiration, lips, tongue, etc. Additionally, this test requires the patient to be with the examiner. It should be noted that both of the evaluations mentioned are subject to clinical criteria and their inter- and intra-rater variability could be high. Computer aided systems could support the clinical diagnosis and improve the evaluation of the disease progression in an objective way. Moreover, these systems could be used by the patient at home, enabling the unobtrusive monitoring of the disease progression and potentially improving the quality of life of the patients.

In recent years, many studies have focused on the detection and evaluation of PD from speech. Two years ago, the INTERSPEECH 2015 Computational Paralinguistics Challenge included the Parkinson's Condition Sub-challenge which addressed the task of predicting the MDS-UPDRS-III scores of PD patients from speech [7]. For this challenge, the recordings of the 50 patients in the PC-GITA database [8] were considered for the train and development subsets. Recordings from 11 different patients were considered as the test set. A neurologic expert assessed the patients with the MDS-UPDRS scale. The motor subscale for each patient was provided to the participants of the challenge, who were asked to predict the labels of the patients in the test set. On the other hand, the Third Frederick Jelinek Memorial Summer Workshop held at Johns Hopkins University in 2016 included Remote Monitoring of Neurodegeneration through Speech as one of the research topics. In [9], the authors used features extracted from speech, handwriting, and gait signals of 30 PD patients. The generalized canonical correlation analysis was used to map these features into other corpora that only included speech. The authors addressed three problems: classification between PD and HC, prediction of the neurological state of the patients according to the MDS-UPDRS-III scale, and prediction of dysarthria level using a modified version of the FDA. The classification accuracy was 78% while the Spearman's correlation between the real and predicted scores was 0.40 for the MDS-UPDRS-III and 0.72 for a modified version of the FDA (m-FDA) which includes 13 items and does not require the patient to be with the examiner.

Speaker adapted models have shown good results in several speech analysis tasks. The speaker-based analysis can be performed using speaker models inspired by the speaker identification and verification fields. A first approach to this kind of analysis can be found in [10]. In that work, the authors used the Gaussian Mixture Model-Universal Background Model (GMM-UBM) approach to model the speech of seven PD patients recorded in four sessions between 2012 and 2015. The aim was to predict the neurological state of each patient

http://dx.doi.org/10.21437/Interspeech.2017-819

through the time. The authors trained an UBM with speech from PD patients and healthy controls (HC). Then, using the speech of one of the sessions of a patient, they adapted the UBM with MAP adaption and obtained a speaker and session specific model. Finally, the Bhattacharyya distance between the speaker and session specific model and the UBM was computed. On average, Pearson's correlation between this distance and the real MDS-UPDRS-III scores was 0.60 . Following this promising approach, in this work we propose to use i-vectors to model the speech of PD patients. This approach was initially proposed for speaker verification in [11] and it has become the state of the art in this and other speech analysis tasks. One of the advantages of this approach is that a simple distance metric can be used for evaluation, namely, the cosine distance. We have two main goals with this study, the first is to check the suitability of i-vectors to model the speech of PD patients and assess their neurological condition. The main hypothesis is that cosine distance between the i-vector of a test speaker and a reference i-vector representing the population of healthy speakers increases if the test speaker suffers from PD. Conversely, the distance with a reference i-vector representing dysarthric speakers will decrease because the test speaker is getting more impaired, thus is getting "closer" to the PD population. Additionally, the i-vectors are built with features related to specific dimensions of speech, e.g., phonation or prosody, to see how much information provides each dimension to represent the speakers. The rest of this paper is organized as follows: Section 2 describes the data and methods. Section 3 describes the experiments and results, and Section 4 includes the conclusions of this study.

2. Methods and materials

The methodology proposed in this study comprises five steps: (1) the speech signals are segmented into voiced/unvoiced segments, (2) features are extracted from the selected segments, (3) a subset of speakers are used to train an i-vector extractor, (4) the i-vectors of every speech signal are extracted and processed as is described in the next paragraph and depicted in Figure 2, and (5) the cosine distance between a reference i-vector and a speaker i-vector is computed. This process is summarized in Figure 1.

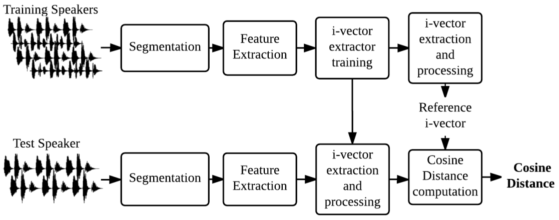
Training Speakers
Segmentation
Feature Extraction
i-vector extractor training
i-vector extraction and processing
Reference i-vector
Test Speaker
Segmentation
Feature Extraction
i-vector extraction and processing
Cosine Distance computation
Cosine Distance

Figure 1: *General methodology followed in this study.*

The i-vectors are processed in five steps: (1) the i-vectors extracted from training speakers are normalized to zero mean and unit variance, i.e., Z-norm, (2) the normalized i-vectors of a given speaker are averaged to obtain one i-vector per speaker, (3) the i-vectors of HC speakers are averaged to obtain the HC reference i-vector, the PD reference i-vector is obtained in a similar way with PD speakers, (4) the i-vectors of a test speaker are Z-normalized using the parameters from the training i-vectors, (5) the normalized i-vectors are averaged to obtain the speaker i-vector. This process is shown in figure 2.

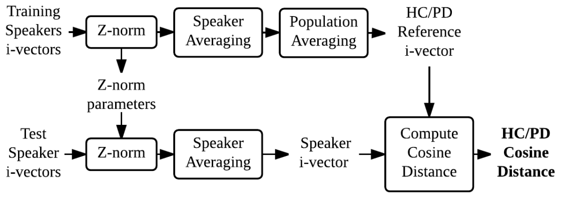

Figure 2: *i-vector processing.*

Once the cosine distances are computed, three analyses are performed: (1) classification between PD and HC, (2) prediction of the neurological state of patients according to the MDS-UPDRS-III score, and (3) prediction of the level of dysarthria according to the m-FDA. Figure 3 depicts these analyses simulating a projection of the i-vectors into a two-dimensional space. As cosine distance compares two vectors in terms of their angle and not their magnitude, the reference and test i-vectors are projected into a unit circle. The angle between the reference and a test i-vector determines the cosine distance. According to our hypothesis, the cosine distance (angle) between a test i-vector and the HC reference i-vector correlates with the speech condition. Classification between PD and HC is performed by comparing this cosine distance with a threshold, and the prediction of the neurological state and the dysarthria level are evaluated with the Spearman's correlation between the cosine distance and the real score.

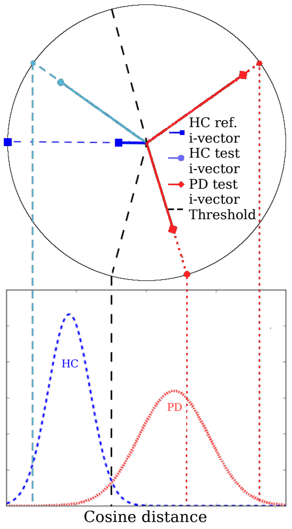

Figure 3: *i-vectors and their projection into Cosine distances.*

2.1. Speech corpus and labels

The PC-GITA speech corpus was used in this study. It contains recordings of 50 PD patients (25 male and 25 female) and 50 healthy controls (HC), all of them native Colombian Spanish speakers. The recordings were captured in a sound-proof booth using professional audio equipment. The original sampling frequency was 44.1 kHz, but the recordings were down-sampled to 16 kHz for this study. During the recordings, the participants were asked to perform different speech tasks including six diadochokinetic (DDK) exercises (e.g., the repetition of /pa-ta-ka/), monologue, read text, and ten short sentences. All the patients were diagnosed by a neurologist expert and assessed according to the MDS-UPDRS [5]. More information of this corpus can be found in [8].

For the m-FDA analysis, three phoniatricians labeled all the subjects in the database. The aim was to evaluate different properties of the speech of patients and controls with more detail than just one item of the MDS-UPDRS-III. The modified version of the FDA is based on speech recordings, thus it does not require the patient to be with examiner. It considers 13 items that evaluate the movements of the lips, larynx and tongue, the respiration, and the intelligibility, among others. The evaluation of each item ranges from 0 to 4 with total range from 0 to 52, with 0 meaning normal and 52 completely dysarthric. The three phoniatricians agreed on the first ten evaluations, and then performed the evaluation of the remaining recordings (inter-rater reliability of 0.75). For this study, each speaker was labeled with the median of the labels given by the three evaluators per speaker.

2.2. Segmentation

Two types of segmentations were performed, one is speech/non-speech which uses an energy-threshold Voice Activity Detector (VAD), and the other one is voiced/unvoiced segmentation, which is done using the autocorrelation method from Praat[12]. A segment is considered voiced when pitch values are detected otherwise it is unvoiced.

2.3. Feature extraction

Four sets of features were considered. The first set comprises the Mel-Frequency Cepstral Coefficients (MFCCs), i.e., the classical features used with the i-vector approach. 19 MFCCs and the log-energy extracted from 30 ms windows with steps of 15 ms were used to form a 20-dimensional feature vector. Non-speech frames were discarded. The other sets are features related to the articulation, phonation, and prosody dimensions of speech. To evaluate articulation, the energy content in the first 22 Bark bands (BBE) in the voiced/unvoiced and unvoiced/voiced transitions were considered as in [13]. The features considered to evaluate phonation and articulation in voiced segments were: the log-energy, pitch (F_0), first and second formants (F_1 and F_2) along with their first and second derivatives, and the Jitter and Shimmer. These form a 14 dimensional vector. The analysis window was 30 ms long and the time shift was 5 ms. To evaluate prosody we followed the approach initially presented in [14]. The log-pitch and log-energy contours within analysis frames were approximated using Lagrange polynomials of order $P = 5$. Analysis frames of 200 ms and a time shift of 50 ms were used [15]. Only the voiced segments within this frames were considered for the approximation. A 13-dimensional feature vector is formed concatenating the 12 coefficients and the number of voiced frames in the segment.

2.4. i-vectors

In this approach, factor analysis is used as a feature extractor to define a new low-dimensional space known as the Total Variability space [11]. This space models the speaker and channel variability. In this case, the speaker variability carries the information about the disorders in speech due the disease. Channel compensation techniques can be applied efficiently in this low-dimensional space, however, such techniques were not used in this work as all the recordings in the speech corpus had the same acoustic conditions, i.e., in this case the Total Variability space only models the speaker variability and the effect of PD in speech. In this approach, an utterance is represented by a supervector of concatenated means and it is expressed as:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where m is an speaker and channel independent supervector (the UBM), T is the Total Variability matrix and w is the i-vector which is a standard normally distributed latent variable.

2.5. Cosine distance

Typically, the cosine kernel (cosine similarity) is used to compare two i-vectors. It is defined as [11]:

$$s_c(w_1, w_2) = \frac{w_1 \cdot w_2}{||w_1||||w_2||}. \tag{2}$$

By normalizing both vectors, this measure considers only the angle between two i-vectors and not their magnitude. The cosine distance is a metric derived from the cosine similarity by taking the positive complement of the latter. This metric can be interpreted as a measure of the dissimilarity between the two vectors. It is defined as:

$$d_c(w_1, w_2) = 1 - s_c(w_1, w_2) \tag{3}$$

In this work, we compare the i-vector of a speaker with a reference i-vector that represents the HC or PD population.

2.6. Evaluation

To classify between healthy controls and PD patients, the cosine distance between the i-vector for the test speaker, w_{spk} and the reference i-vector is compared with a threshold θ. For instance, for the HC reference i-vector, w_{HC}, the threshold analysis would be expressed as: $d_c(w_{spk}, w_{HC}) \geq \theta$. By varying θ it is possible to obtain the False Positive Rate (FPR) and False Negative Rate (FNR), which also allows the computation of the Equal Error Rate (EER) and the Area Under the ROC Curve (AUC) [16]. Results for this analysis will be given in terms of the accuracy (Acc.) at the EER and the AUC. The tasks of predicting the neurological state of a patient and the dysarthria level are evaluated using the Spearman's correlation coefficient (ρ).

3. Experiments and results

Each speech task described in the previous section is analyzed independently. To obtain cosine scores for each speaker in the database, ten random divisions are formed. Each random division has 90 speakers for training (45 PD patients and 45 HC) and 10 speakers for testing (5 PD patients and 5 HC). In each division the test speakers are always different and speaker independence is guaranteed in all of the process. The 90 speakers in each training set were used in several processes: (i) to train the UBM and the i-vector extractor, (ii) to compute the z-normalization parameters, and (iii) to obtain the HC and PD reference i-vectors. UBMs with different number of Gaussians were tested varying from $M = 2$ to $M = 2^9$ in powers of 2. The dimension of the i-vector, $\dim_w$, was chosen based on the number of Gaussians in the UBM and the dimension of the original feature vector, $\dim_f$, following the relation $\dim_w = \log_2(M)\dim_f$. The speech signals of each test speaker were used to obtain their speaker specific i-vectors. The cosine distances between each speaker specific i-vector and both HC and PD reference i-vectors were computed. Once the cosine distances for the 100 speakers are computed, classification and correlation experiments are performed. Due to space limitations, only results of comparing the threshold with the HC reference i-vector are presented. Very similar results were found using the PD reference i-vector. Only the patients's cosine distances were considered to compute the correlation with the MDS-UPDRS-III labels because there are no labels for the HC. When computing the correlation with the m-FDA, both patient's and healthy speakers are considered as we have the m-FDA labels for all of the 100 speakers in PC-GITA. The modeling and extraction of the i-vectors were done using Kaldi [17]. The rest of the processes were performed using Python.

Table 1: *Classification results in terms of Acc. and F1 score at EER and AUC*

	DDK			Monologue		
Features	Acc.	F1	AUC	Acc.	F1	AUC
MFCCs	76%	0.76	**0.85**	**78%**	**0.78**	**0.85**
Art.	74%	0.74	0.82	76%	0.76	0.83
Phon./Art.	**78%**	**0.78**	0.84	72%	0.72	0.81
Prosody	74%	0.74	0.80	66%	0.66	0.65
	Read text			Sent.		
Features	Acc.	F1	AUC	Acc.	F1	AUC
MFCCs	74%	0.74	0.83	**78%**	**0.78**	**0.90**
Art.	74%	0.74	**0.84**	**78%**	**0.78**	0.88
Phon./Art.	70%	0.70	0.80	74%	0.74	0.80
Prosody	64%	0.64	0.68	60%	0.60	0.69

Table 1 presents the classification results in terms of AUC. The best result is obtained using the classical MFCCs-based approach in sentences, while the best accuracy is obtained with the articulation features measured upon the DDK task. Phonation features only perform well in the DDK speech task. The F1 scores are the same as the corresponding accuracies, this is due to the balance between patients and HC in the database. Table 2 shows the correlation values between the cosine distance to the reference i-vectors and the real MDS-UPDRS-III scores. The best results are obtained with the i-vectors formed with the phonation features extracted from sentences. Table 3 shows the results of the correlation between the cosine distance to the reference i-vectors and the m-FDA. Note that both phonation and prosody features exhibit the best performance in the DDK task.

Note that in all of the cases the correlations with the HC reference i-vector are positive while the correlations with the PD reference i-vector are negative. This behavior indicates that PD i-vectors and HC i-vectors are opposed in the “i-vector space”. As this is systematically obtained in our experiments, this is a very promising result to represent healthy speakers and people with speech disorders. Additionally, it is important to highlight that the correlation values are computed here between two one-dimensional vectors, thus negative values imply an opposite trend between those vectors.

4. Conclusion

This work evaluated the suitability of i-vectors to model the speech of Parkinson's patients and assess their condition. The i-vectors extracted from training speech signals were used to create a reference i-vector that represented the population of HC or PD patients. The cosine distance between a test speaker i-vector and a reference i-vector was computed. This distance was used in three different experiments: (i) to classify between PD and HC, (ii) to assess the neurological state of the patient according to the MDS-UPDRS-III scale, and (iii) to predict a modified version of the Frenchay Dysarthria Assessment. The results show that this approach is suitable to evaluate the speech of people with Parkinson's disease. A positive correlation between the labels and the cosine distance to the HC reference i-vector was found. This was expected, because the more affected the speech, the larger the distance to the healthy speakers. Similarly, negative correlations were found when comparing test speakers w.r.t. the PD reference i-vector, i.e., the more affected the speech, the lower the distance to the PD speakers. Mel-Frequency Cepstral Coefficients were compared to feature sets that provide information about three dimensions of speech: articulation, phonation, and prosody. In many cases phonation and prosody features exhibited better results than MFCCs. This confirms that considering certain speech dimensions independently improves the performance of the evaluation and allows further interpretation [18]. Future work includes the use of these features to predict different sub-scores of the m-FDA to do further interpretations, and the use of i-vectors to assess the disease progression per speaker like in [10].

Table 2: *Spearman's correlation between the MDS-UPDRS-III labels and the cosine distance with the reference i-vectors*

– with respect to the reference HC i-vector –				
Features	DDK	Monologe	Read text	Sent.
MFCCs	0.38	**0.54**	0.25	0.49
Articulation	0.35	0.49	0.30	0.45
Phon. & Art.	**0.48**	0.34	**0.41**	**0.63**
Prosody	0.27	0.34	0.28	0.39
– with respect to the reference PD i-vector –				
Features	DDK	Monologe	Read text	Sent.
MFCCs	-0.38	**-0.54**	-0.25	-0.49
Articulation	-0.34	-0.49	-0.30	-0.45
Phon. & Art.	**-0.48**	-0.34	-0.41	**-0.63**
Prosody	-0.27	-0.34	-0.28	-0.39

Table 3: *Spearman's correlation between the m-FDA labels and the cosine distance with the reference i-vectors*

– with respect to the reference HC i-vector –				
Features	DDK	Monologue	Read text	Sent.
MFCCs	0.57	**0.63**	0.49	**0.64**
Articulation	0.54	0.54	**0.54**	0.62
Phon. & Art.	**0.72**	0.50	0.46	0.56
Prosody	**0.72**	0.27	0.34	0.36
– with respect to the reference PD i-vector –				
Features	DDK	Monologue	Read text	Sent.
MFCCs	-0.57	**-0.63**	-0.49	**-0.64**
Articulation	-0.54	-0.54	**-0.54**	-0.62
Phon. & Art.	**-0.72**	-0.50	-0.46	-0.56
Prosody	**-0.72**	-0.27	-0.34	-0.36

5. Acknowledgements

This work was partially funded by COLCIENCIAS through the project No 111556933858 and CODI through project PRV16-2-01. It was partially developed during the JSALT 2016 supported by JHU via grants from DARPA (LORELEI), Microsoft, Amazon, Google, and Facebook. N. Garcia was granted with the SIPGA scholarship by A*STAR and I^{2}R.

6. References

[1] S. Sveinbjornsdottir, “The clinical symptoms of Parkinson’s disease,” *Journal of Neurochemistry*, no. 139, pp. 318–324, jul 2016.

[2] A. A. Moustafa *et al.*, “Motor symptoms in Parkinson’s disease: A unified framework,” *Neuroscience and Biobehavioral Reviews*, vol. 68, pp. 727–740, 2016.

[3] J. A. Stamford *et al.*, “What engineering technology could do for quality of life in Parkinson’s disease: A review of current needs and opportunities,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1862–1872, nov 2015.

[4] K. Tjaden, “Speech and Swallowing in Parkinson’s Disease.” *Topics in geriatric rehabilitation*, vol. 24, no. 2, pp. 115–126, 2008.

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

[6] P. Enderby and R. Palmer, *FDA-2: Frenchay Dysarthria Assessment*, 2nd ed. P. Education, 2008.

[7] B. Schuller *et al.*, “The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson’s & eating condition,” in *Proceedings of the 16th NTERSPEECH*, 2015, pp. 478–482.

[8] J. R. Orozco *et al.*, “New Spanish speech corpus database for the analysis of people suffering from Parkinson’s disease,” in *Proceedings of the 9th LREC*, 2014, pp. 342–347.

[9] J. C. Vasquez-Correa *et al.*, “Multi-view Representation Learning Via GCCA for Multimodal Analysis of Parkinson’s Disease,” in *Proceedings of the 42nd ICASSP*, 2017.

[10] T. Arias-Vergara *et al.*, “Parkinson’s Disease Progression Assessment from Speech Using GMM-UBM,” in *Proceedings of the 17th NTERSPEECH*, 2016, pp. 1933–1937.

[11] N. Dehak *et al.*, “Front-End Factor Analysis for Speaker Verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, may 2011.

[12] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193, pp. 97–110, 1993.

[13] J. Orozco-Arroyave *et al.*, “Towards an automatic monitoring of the neurological state of Parkinson’s patients from speech,” in *Proceedings of the 41st ICASSP*, 2016, pp. 6490–6494.

[14] N. Dehak *et al.*, “Modeling prosodic features with joint factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[15] D. Martínez *et al.*, “iVector-based prosodic system for language identification,” in *Proceedings of the 37th ICASSP*, mar 2012, pp. 4861–4864.

[16] N. Sáenz-Lechón *et al.*, “Methodological issues in the development of automatic systems for voice pathology detection,” *Biomedical Signal Processing and Control*, vol. 1, pp. 120–128, 2006.

[17] D. Povey *et al.*, “The kaldi speech recognition toolkit,” in *Proceedings of the IEEE ASRU*, 2011.

[18] J. Orozco-Arroyave, *Analysis of speech of people with Parkinson’s disease*, 1st ed. Berlin, Germany: Logos-Verlag, 2016.