EXPLORING THE USE OF PHONOLOGICAL FEATURES FOR PARKINSON'S DISEASE DETECTION

Nina Hosseini-Kivanani [1], Juan Camilo Vásquez-Correa[2], Christoph Schommer[1], Elmar Nöth[3]

[1]Faculty of Science, Technology, and Medicine, University of Luxembourg, [2]Vicomtech Foundation,

Basque Research and Technology Alliance (BRTA) Donostia-San Sebastian, Spain,[3]Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nüremberg, Germanyy

nina.hosseinikivanani@uni.lu, jcvasquez@vicomtech.org, christoph.schommer@uni.lu, elmar.noeth@fau.de

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder that causes motor and non-motor symptoms. Speech impairments are one of the early symptoms of PD, but they are not always fully exploited by clinicians. In this study, the use of phonological features extracted from speech data collected from Spanish-speaking patients was explored to predict PD patients from healthy subjects using phonet, which was trained on Spanish data, and PhonVoc, which was trained on English data. These features were then used to train and test several machine learning models. The XGBoost model achieved the best performance in classifying patients from HCs, with an accuracy of over 0.76. However, the model performed better when using a phonological model trained on Spanish data rather than English data.

Keywords: Parkinson's disease, machine learning models, classification, Phonet, PhonVoc.

1. INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's that affects movement and can cause symptoms such as tremor, slowness of movement, muscle rigidity, and different writing and speech deficits. PD is typically diagnosed through a combination of a patient's medical history, symptoms, and a neurological and physical exam. However, PD is a slow-progressive disease and may have a prodromal period of 3 to 15 years during which the main motor symptoms may not be clearly perceivable [1]. This highlights the importance of developing robust and automatic methods for detecting the disease in its early stages to improve the patient's quality of life.

Speech impairments are often one of the earliest motor symptoms of PD, and they can worsen as the disease progresses. Recent studies have indicated that over 90% of PD patients express some kind of speech impairment [2], with disorders mostly related to phonation and articulation, including alterations in speech rate and pitch variations, imprecise articulation of vowels and consonants, and monotonous speech, leading to decreased speech intelligibility [3]. Speech impairments in individuals with PD often result in imprecise articulation, particularly affecting the production of stop, affricate, and fricative sounds [4]. Analysis of speech materials such as sustained vowels and diadochoknetic (DDK) exercises (i.e., rapid syllable repetition) can aid in monitoring the disease severity of patients [5]. which have been widely used for detecting early symptoms of the disease [6].

Various machine learning (ML) methods have been widely used to classify PD patients vs. HCs using speech signals ([7, 8]). In [7], feature subset selection ranging from 8 to 20 was used to represent dimensionality reduction in complexity, and the accuracy of ML models was improved using 10-fold cross-validation on a small dataset to implement a voice-based detection methodology In [9], the importance and effectiveness of ML models were investigated using utterances of the sustained vowel /ah/ from 188 PD patients and 64 HCs. The results showed that repeating voice patterns are identified in PD speech, proving that ML models can support the diagnosis of PD with similar accuracy as movement disorder therapists.

Phonological posterior features have been found to be helpful in gathering information about the presence and severity of PD in patients [10]. Among the common toolkits for extracting phonological features, PhonVoc [11] has been used to evaluate the progress of apraxia based on speech. [12] used PhonVoc to extract phonological features of impaired speech to better distinguish the disease. In addition, the Phonet toolkit [13] is also being used for speech processing, specifically for predicting the posterior probability of speech files, which refers to the likelihood of a speech file based on the data it contains. [14] recently used Phonet to extract phonological features from Dutch patients with dysarthria to develop a computer-based therapy approach. This study aims to address the pertinent problem of identifying and monitoring PD using speech signals as a measurement, with the intention of creating useful tools for this purpose.

It is uncertain how neurodegenerative diseases impact the production of different groups of phonemes, including those related to the manner and place of articulation. Additionally, the majority of available language resources are in English. We are conducting a study to compare the performance of two phonological feature extraction models, one trained on English data and one trained on Spanish data. This is the first cross-language study of these models, and the goal is to determine whether these models are language independent or if one model performs better on the Spanish data due to being trained on data in the same language.

2. MATERIAL & METHODS

2.1. PD Dataset

The study used data from the PC-GITA database [15], which consists of speech utterances from 68 individuals with PD and 48 healthy controls (HC), all of whom were Spanish-speaking and balanced in terms of age and gender. Patients' neurological state was evaluated using the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS-III scale) [16]. Recording were taken while patients were on anti-Parkinsonian medication. This was done because of ethical reasons. The statistical tests for validating gender and age balance are detailed in Table 1([17]).

2.1.1. Speech material

The Speech material includes six DDK tasks, which consist of rapid repetition of syllables such as /pa-ta-ka/, /pa-ka-ta/, /pe-ta-ka/, /pa/, /ta/, and /ka/. These tasks are designed to evaluate the speakers' ability to move their vocal tract articulators (e.g, lips, tongue, and soft palate) as quickly and consistently as possible.

2.2. Toolkits

We used **PhonVoc** [11] and **Phonet** [13] toolkits to extract phonological features from the speech signal, including features related to phonation. **PhonVoc** is a toolkit that uses a deep neural network (DNN) to estimate the probability that a sound belongs to a specific phonological class in English, and it has an accuracy of over 96% for detecting these classes. **Phonet** toolkit is based on Gated Recurrent Neural Networks to extract phonological posteriors from speech. The toolkit was trained on 17 hours of clean FM podcasts in Mexican Spanish, which is able to detect the phonological classes with high accuracy (90%). For each toolkit, we have calculated values for the mean, standard deviation, skewness, and kurtosis. The list of phonological classes that correspond to both stop and voiced segments of speech for evaluating consonant imprecision includes "Back," "Voice," "Vocalic," "Consonantal," "Continuant," "Coronal," & "Silence". "Silence" is considered a separate phonological class, because it represents the absence of sound. These classes can be used to analyze the accuracy of the toolkits in detecting specific phonological features in speech signals.

2.3. Classical ML models

In this project, ML techniques were used to classify speech based on its features. ML allows computers to recognize patterns in data and make decisions based on those patterns. We used several commonly used ML models such as Naïve Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), *k*-nearest neighbors (*k*NN), and Ensemble methods (Voting & eXtreme Gradient Boosting (XGBoost)) for the binary classification problem. These models were implemented using the *scikit-learn* library in python (3.9).

2.3.1. Data sampling methods

The number of speech samples for HCs and PD patients was not equal: 72% of the samples related to tasks performed by PD patients, while 28% related to tasks performed by HCs. Imbalanced distribution of the classes could lead to a higher likelihood of misclassification for the minority class. A balancing technique was applied to create equal samples for each class. We used a number of commonly employed techniques for handling imbalanced datasets, including:

Oversampling: It is a technique for balancing imbalanced datasets by increasing the number of minority class samples in the training dataset. One commonly used oversampling technique is Synthetic Minority Over-sampling (SMOTE) [18], which generates synthetic data by selecting the nearest data points (typically k = 5) based on the

	PD patients	HC subjects	Patients vs. controls
Gender [F/M]	31/37	24/24	$^{*}p = 0.99$
Age [F/M]	60.9(11.2)/64.7(9.4)	59.9(8.7)/63.5(10.4)	$^{**}p = 0.08$
Time since diagnosis [F/M]	15.5(14.5)/8.1(5.9)	–	
MDS–UPDRS–III [F/M]	36.2(18.1)/36.3(18.9)	–	

Time since diagnosis and age are given in years. [F/M]: Female/Male. Average (Standard deviation).

$^{*}p$–value calculated through Chi–square test. $^{**}p$–value calculated through Mann-Whitney U-test.

Table 1: Clinical & demographic information of the subjects from the PC-GITA database [15].

Euclidean distance between them in the feature space and duplicating these points in the dataset.

Class weight: This technique involves computing the frequency of each class and then inverting it such that the underrepresented class has a much larger error when multiplied by the class loss compared to the majority class.

Threshold: It is a technique for adjusting the predicted values by setting a threshold value. If the predicted value is greater than this threshold, it is set to 1; otherwise, it is set to 0.

2.4. Evaluation

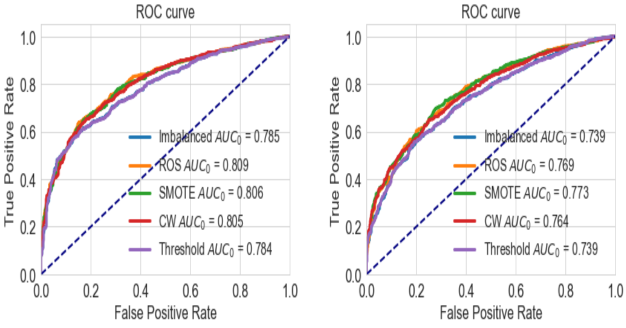
The dataset was split into a training set (75% of data, including 84 out of 112 participants) and a testing set (25% of data) based on the participants' IDs. To optimize performance, k-fold cross validation was used, specifically **group k-fold** cross validation (n_splits=5), to ensure the same group of recordings from one participant does not appear in both the training and testing sets, avoiding overfitting and ensuring subject-independent cross-validation. To evaluate the performance of the ML models, we used metrics such as accuracy, precision, recall, and receiver operating characteristic (ROC) to obtain the Area Under Curve (AUC). These metrics are based on four outcomes: true negative (TN), false negative (FN), false positive (FP), and true positive (TP), which are shown in a confusion matrix (Figure 3).

- Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$
- AUC $= \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$

3. RESULTS & DISCUSSION

The outputs of using Phonet to extract Spanish phonological features (SpNet) and English phonological features (EnVoc) were presented to evaluate the performance of the models with these features. By comparing the results with different types of features, it is possible to determine which models work best with specific types of data. To evaluate the effectiveness of balancing techniques, we used SVM for binary classification tasks which yields the best accuracy score with relatively low training times. We calculated the accuracy scores of the models with and without a balanced dataset for the baseline model to determine which balancing techniques were most effective.

The baseline classifier model had an accuracy of 0.75 for the SpNet and 0.73 for the EnVoc, but it was biased towards the majority class (PD). To address this issue, various balancing approaches were employed (as described in Section 2.3.1). The results of the baseline model on different sampling techniques (Figure 1) showed that oversampling performed better than SMOTE, class weight, or threshold techniques, with AUC values of 0.81 and 0.77 for SpNet and EnVoc, respectively. AUC values between 0.7 and 0.8 are considered fair [19] , and the values obtained for SpNet and EnVoc fall within this range. This suggests that the model is relatively accurate in its classification tasks.

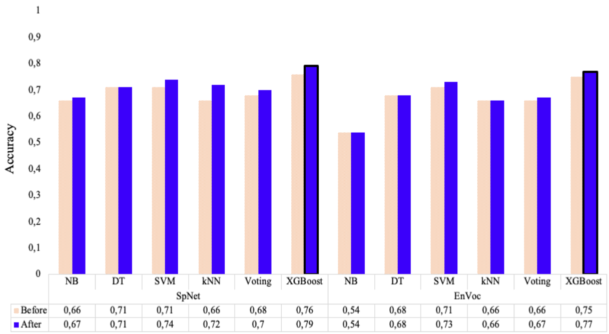


(a) SpNet

(b) EnVoc

Figure 1: ROC curves & AUC values: imbalanced, Oversampling, SMOTE, Class weight & threshold.

To further improve the performance of the ML models, we implemented a grid search procedure to optimize the hyperparameters of the models. This entailed specifying a range of values for each hyperparameter and evaluating the performance of all possible combinations to identify the optimal configuration. Although this did not significantly improve as a result of this process, with the exception of the SVM model, the results of the other models were still relatively comparable to those of the SVM model.



	SpNet						EnVoc					
	NB	DT	SVM	kNN	Voting	XGBoost	NB	DT	SVM	kNN	Voting	XGBoost
Before	0,66	0,71	0,71	0,66	0,68	0,76	0,54	0,68	0,71	0,66	0,66	0,75
After	0,67	0,71	0,74	0,72	0,7	0,79	0,54	0,68	0,73	0,66	0,67	0,77

Figure 2: Result of ML models before & after hyperparameter tuning.

In addition to the ML models previously mentioned, we also used ensemble methods such as the Voting method and the XGBoosting technique to try to improve the performance of the ML models. The Voting method involves combining the results of multiple base classifiers based on weighting, while XGBoosting is a combination of bagging and boosting techniques. After applying hyperparameter tuning and oversampling to the datasets, we found that the XGBoost model was more effective at classifying PD vs. HC than the other models, achieving a mean accuracy of 0.79 for SpNet and 0.77 for EnVoc 2. The confusion matrix for this model also supported its effectiveness (Figure 3).

The confusion matrix in Figure 3 illustrates the model's performance and the types of errors it is making by comparing the actual labels (vertical levels) to predicted labels (horizontal levels) by XGBoost model for each phonological feature to evaluate the model's performance in the binary classification problem. These values were used to calculate accuracy, recall, and precision. The results of the XGBoost model are presented in Figure 3, which shows the correct (i.e., TP and TN) and incorrect (i.e., FP and FN) predictions made by the model. These results demonstrate high recall and precision values: recall for SpNet was 0.85 and for EnVoc was 0.84, while precision for SpNet was 0.84 and for EnVoc was 0.83.

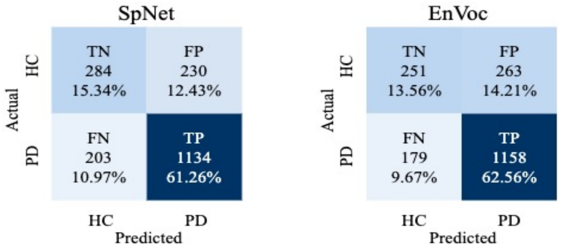


Figure 3: Confusion matrix of XGBoost model.

The XGBoost model was used to conduct a feature importance analysis in PD speech by ranking the input features based on their phonological features on SpNet and EnVoc. This analysis identified the most important and influential features for classifying PD speech from HCs in the dataset. The top team important features are shown in Figure 4, with higher scores indicating a greater impact on the model's predictions. The mean of continuant in SpNet and the kurtosis of voice consonant in EnVoc received the highest importance scores among all the features (Figure 4). These listed features have the potential to improve the model's performance. The underlying aspects of speech production warrant further investigation.

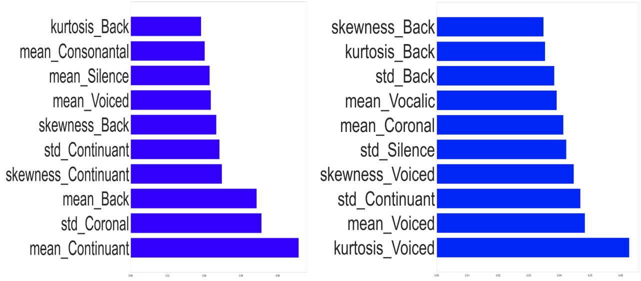


Figure 4: Top 10 important features of SpNet (left) & EnVoc (right) for XGBoost model.

4. CONCLUSION

We conducted a study to assess the ability of ML models to classify PD patients and HCs based on speech features. We compared the performance of phonVoc and phonet toolkits: one trained on a large corpus of Librispeech consisting of 1000 hours of clear speech in English, and another trained on the Spanish corpus of radio podcasts in Spanish. Our results showed that SpNet had slightly better performance than English (0.79 vs 0.77). The difference in the quality and size of the training data may have contributed to the slightly lower performance of phonVoc model. But this difference could be larger if we had access to a higher quality corpus of Spanish speech. Our study is the first to compare the performance of ML models trained on English and Spanish for our task, and this opens up new research avenues, such as considering the use of the UPDRS-III for multiclass classification problems. These results support the usefulness of the Phonet toolkit, which was specifically designed for the Spanish language and optimized for classifying speech data.

5. ACKNOWLEDGEMENT

This publication is based upon work from COST Action GoodBrother, COST Action

19121—Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living, supported by COST (European Cooperation in Science and Technology).

6. REFERENCES

[1] R. Savica, J. M. Carlin, B. R. Grossardt, J. H. Bower, J. E. Ahlskog, D. M. Maraganore, A. E. Bharucha, and W. A. Rocca, "Medical records documentation of constipation preceding Parkinson disease," *Neurology*, vol. 73, no. 21, pp. 1752–1758, 2009.

[2] G. Solana-Lavalle, J. C. Galán-Hernández, and 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*, vol. 40, no. 1, pp. 505–516, 1 2020.

[3] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Scientific Reports*, vol. 7, no. 1, p. 12, 2017.

[4] J. A. Logemann and H. B. Fisher, "Vocal Tract Control in Parkinson's Disease," *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, pp. 348–352, 1981.

[5] J. C. Vásquez-Correa, C. D. Rios-Urrego, A. Rueda, J. R. Orozco-Arroyave, S. Krishnan, and E. Nöth, "Articulation and Empirical Mode Decomposition Features in Diadochokinetic Exercises for the Speech Assessment of Parkinson's Disease Patients," in *Iberoamerican Congress on Pattern Recognition*, vol. 11896 LNCS. Springer, 2019, pp. 688–696.

[6] F. Karlsson, E. Schalling, K. Laakso, K. Johansson, and L. Hartelius, "Assessment of speech impairment in patients with Parkinson's disease from acoustic quantifications of oral diadochokinetic sequences," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, p. 839, 2 2020.

[7] G. Solana-Lavalle and R. Rosas-Romero, "Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation," *Biomedical Signal Processing and Control*, vol. 66, p. 102415, 2021.

[8] J. R. Orozco-Arroyave, J. C. Vádsquez-Correa, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, no. March, pp. 6490–6494, 2016.

[9] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters*, vol. 125, pp. 55–62, 2019.

[10] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vásquez-Correa, and E. Nöth, "Characterisation of voice quality of Parkinson's disease using differential phonological posterior features," *Computer Speech and Language*, vol. 46, no. June, pp. 196–208, 2017.

[11] M. Cernak and P. N. Garner, "PhonVoc: A phonetic and phonological vocoding toolkit," in *INTERSPEECH*, 2016, pp. 988–992.

[12] A. Asaei, M. Cernak, and M. Laganaro, "PAoS Markers: Trajectory Analysis of Selective Phonological Posteriors for Assessment of Progressive Apraxia of Speech," in *Proceeding on the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016.

[13] J. C. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, "Phonet: a Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech." in *INTERSPEECH*, 2019, pp. 549–553.

[14] V. M. Ramos, J. C. Vasquez-Correa, R. Cremers, L. Van, D. Steen, E. Nöth, M. De Bodt, and G. V. Nuffelen, "Automatic boost articulation therapy in adults with dysarthria: Acceptability, usability and user interaction," *International Journal of Language & Communication Disorders*, vol. 56, no. 5, pp. 892–906, 2021.

[15] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease." in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 342–347.

[16] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, and others, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[17] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease," *Journal of Communication Disorders*, vol. 76, pp. 21–36, 2018.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[19] J. V. Carter, J. Pan, S. N. Rai, and S. Galandiuk, "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves," *Surgery*, vol. 159, no. 6, pp. 1638–1645, 2016.