Project Proposal

● Ungraded

Student

Thu Vu

✎ View or edit group

Total Points

- / 1 pts

Question 1

Project Proposal 1 pt

Question assigned to the following page: 1

Speech-Based Early Biomarker Identification and Classification for Parkinson’s Disease

Authors: Raksha Bellary , Xirui Liu , Thu Vu , Yinan Zhu

I. Introduction

Parkinson’s disease (PD) is a neurological disorder that causes weakening and damage in neurons, ultimately leading to movement problems such as tremor, stiffness, and inability to balance (NIH) [1] Due to the lack of a cure for PD, early detection is crucial for disease management and treatment. Hence, we seek to use machine learning to improve early diagnosis and provide further insight into PD progression. In this project, a dataset consisting of speech features from patients with early untreated PD, patients at high risk of PD, and healthy controls is leveraged to identify key features that may allow for more accurate classification of PD, primarily during early stages.

II.Related Works

>>> Original Paper: Hlavnička, J., Čmejla, R., Tykalová, T. et al. Automated analysis of connected speech reveals early biomarkers of Parkinson’s disease in patients with rapid eye movement sleep behaviour disorder. Sci Rep 7, 12 (2017).

The study introduces a fully automated method for analyzing speech abnormalities in neurodegenerative disorders like PD using acoustic data from natural speech. Traditional evaluations relied on perceptual tests or small-scale analyses, but this approach identifies key speech features linked to respiratory deficits, dysphonia, imprecise articulation, and dysrhythmia. Comparing speech recordings from 50 RBD patients, 30 newly diagnosed PD patients, and 50 healthy controls, the study found that early parkinsonian speech deficits can be detected even in RBD patients, who are at high risk of developing PD. Their findings suggest that automated vocal analysis could enhance early screening and diagnosis of PD in real-world settings [2].

>>> Another supporting paper: Tracy, John M., et al. "Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease." *Journal of biomedical informatics* 104 (2020): 103362.

The study explores the use of voice as a biomarker for early detection of Parkinson’s disease (PD) through machine learning-based deep phenotyping methods. Researchers utilized the *mPower* dataset, which includes voice recordings of PD patients and healthy controls, to extract paralinguistic features for analysis. Feature extraction involved using MATLAB’s Voicebox toolkit and OpenSmile to derive 2268 features related to speech articulation, phonation, and frequency modulation. Three conventional machine learning

Question assigned to the following page: 1

models, namely, the L2-regularized logistic regression, random forest, and gradient-boosted decision trees, were trained on the dataset to classify early-stage PD from controls. The gradient-boosted decision trees model outperformed the others, achieving a recall of 0.646, precision of 0.849, F1-score of 0.688, and an AUC of 0.88. However, the study identified identity confounding, where multiple voice recordings from the same participants led to overestimated performance. To mitigate this, a revised train-test split ensured that no individual appeared in both training and testing datasets, which reduced performance but still maintained strong classification capabilities. Key acoustic features distinguishing PD from healthy controls included fundamental frequency variability (F0), spectral slope, and formant frequency shifts. The findings suggest that voice biomarkers can be a non-invasive, cost-effective tool for PD screening and progression monitoring. However, the study emphasizes the need for larger, clinically validated datasets and longitudinal studies to confirm the robustness of these models in real-world applications [3].

III. Methods

1. Feature Selection: We plan to utilize feature selection in order to reduce the dimensionality of our feature space and find the key features that contribute to high performance in early PD detection. Simultaneously, we will be able to reduce the noise in our data to help us achieve a more refined classifier. Specifically, one method of feature selection that we plan to use is Principal Component Analysis (PCA).

2. Classification using Support Vector Machines (SVM): We plan to preprocess the dataset by normalizing speech features to ensure consistent scaling. We will then split the data into training and testing sets to ensure balanced representation of early PD, high-risk individuals, and healthy controls. An SVM with a radial basis function (RBF) kernel will be trained to find an optimal decision boundary that separates the different groups based on extracted acoustic features. Feature selection techniques like Recursive Feature Elimination (RFE) will be applied to retain the most relevant speech biomarkers. Model performance will be evaluated using accuracy, precision, recall, F1-score, and AUC-ROC, ensuring robustness in classification. Additionally, hyperparameter tuning via grid search will optimize the C (regularization) and gamma (kernel coefficient) parameters to improve model generalization.

Question assigned to the following page: 1

3. Deep Learning with convolutional neural network (CNN) where we test whether the network structure can capture complex speech patterns by extracting spectro-temporal features, ultimately improving classification accuracy.

IV. Performance Evaluation

- Performance of feature selection will be evaluated by comparing the classifier's performance on original and reduced feature space.
- Performance of the classification models will be evaluated using the following four metrics: accuracy, precision, recall, F1-score.

V.References

[1] National Institute of Neurological Disorders and Stroke. Parkinson's Disease. U.S. Department of Health and Human Services, https://www.ninds.nih.gov/health-information/disorders/parkinsons-disease#:~:text=What%20is%20Parkinson%27s%20disease?,or%20completing%20other%20simple%20tasks. Access March 14th, 2025.

[2] Hlavnička, J., Čmejla, R., Tykalová, T. et al. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. Sci Rep 7, 12 (2017).

[3] Tracy, John M., et al. "Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease." *Journal of biomedical informatics* 104 (2020): 103362.