Analyzing Voice Measures as Predictors of Parkinson's Disease Progression

Madeline Dabney

Abstract

This study investigates the potential of voice measures as biomarkers for monitoring Parkinson's disease (PD) progression. Utilizing the Oxford Parkinson's Telemonitoring dataset, we analyzed 5,875 voice recordings from 42 individuals with early-stage PD over six months. Our analysis employed Ordinary Least Squares regression to explore the relationship between various vocal features and Unified Parkinson's Disease Rating Scale (UPDRS) motor scores. The regression model explained 15% of the variance in motor UPDRS scores, with significant predictors including age, sex, and voice parameters such as Jitter(%), Jitter(Abs), and Pitch Period Entropy (PPE). Principal Component Analysis revealed that the first component accounted for 66.2% of the dataset variance, suggesting potential for dimensionality reduction in future models. While our findings support the use of voice analysis as a non-invasive method for tracking PD progression, the moderate predictive power indicates that it should be used in conjunction with other clinical assessments. This study contributes to the growing body of evidence supporting voice-based PD monitoring, potentially enabling more frequent and accessible patient evaluation. Future research should focus on longitudinal studies, larger datasets, and advanced machine learning techniques to improve the predictive accuracy of voice-based PD assessment.

Keywords: Parkinson's Disease, Voice Measures, UPDRS, Biomarkers, Telemonitoring

1 Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that significantly impacts patients' quality of life, affecting movement, balance, and speech. It is characterized by the degeneration of dopaminergic neurons in the substantia nigra, leading to motor symptoms such as bradykinesia, rigidity, and tremors. Early detection and accurate monitoring of disease progression are crucial for effective management and intervention strategies (Tsanas et al., 2010).

Recent studies have demonstrated that changes in vocal characteristics can reflect underlying motor dysfunctions associated with PD, potentially serving as biomarkers for disease progression. Voice analysis offers a non-invasive and cost-effective method for assessing PD symptoms, making it an attractive option for frequent patient monitoring. Various acoustic measures, such as jitter, shimmer, and harmonics-to-noise ratio (HNR), have been found to correlate with PD severity and progression (Frid et al., 2021; Mei et al., 2022). These voice features can be objectively measured and analyzed, providing quantitative data on subtle changes in vocal performance that may not be apparent in clinical examinations.

This study aims to investigate the relationship between biomedical voice measures and clinical indicators of PD progression, specifically the motor and total UPDRS (Unified Parkinson's Disease Rating Scale) scores. By analyzing the Oxford Parkinson's Telemonitoring dataset, which comprises 5,875 voice recordings from 42 individuals with early-stage PD collected over six months, we seek to identify potential vocal biomarkers for disease progression (Tsanas

& Little, 2009). This dataset provides a valuable resource for exploring the correlation of various voice characteristics with disease severity.

Understanding the correlation between voice characteristics and PD progression could lead to improved monitoring techniques and more timely therapeutic interventions. Recent advancements in machine learning have shown promise in classifying PD based on voice recordings, achieving high accuracy rates (Little et al., 2009). This research contributes to the growing body of evidence supporting the use of voice analysis in PD assessment and management, potentially enabling more frequent and accessible patient evaluation in both clinical and home settings.

2 Data Description

This study utilizes the Oxford Parkinson's Telemonitoring dataset, obtained from the UCI Machine Learning Repository[6]. The dataset comprises 5,875 voice recordings collected from 42 individuals with early-stage Parkinson's disease over a six-month clinical trial of a telemonitoring device for remote symptom progression monitoring.

2.1 Dataset Overview

Characteristic	Value
Number of recordings	5,875
Number of subjects	42
Recordings per patient (approx.)	200

Table 1: Dataset Overview

2.2 Variables

The dataset contains 22 variables, including:

- Demographic information: subject number, age, and sex
- Clinical scores: motor UPDRS and total UPDRS
- Voice measures: 16 biomedical voice features

2.3 Demographic and Clinical Characteristics

Characteristic	Mean	Std Dev	Min	Max
Age (years)	64.80	8.82	36	85
Motor UPDRS	21.30	8.13	5.04	39.51
Total UPDRS	29.02	10.70	7.00	54.99

Table 2: Demographic and Clinical Characteristics

Sex distribution: 68.2% male, 31.8% female

2.4 Voice Measures

The dataset includes various voice measures that capture different aspects of vocal impairment associated with Parkinson's disease:

- Jitter: Measures of frequency instability
- Shimmer: Measures of amplitude instability
- NHR and HNR: Indicators of the degree of hoarseness
- RPDE (Recurrence Period Density Entropy): A nonlinear measure of periodicity
- DFA (Detrended Fluctuation Analysis): A nonlinear measure of stochastic self-similarity
- PPE (Pitch Period Entropy): A nonlinear measure of impairment in pitch period

This comprehensive set of voice measures allows for a detailed analysis of the relationship between vocal characteristics and Parkinson's disease progression as indicated by the UPDRS scores.

3 Methods

This study employs a quantitative analysis approach to investigate the relationship between biomedical voice measures and clinical indicators of Parkinson's disease (PD) progression. The analysis was conducted using Python 3.11, leveraging several libraries including Pandas for data manipulation, NumPy for numerical computations, and Statsmodels for statistical modeling.

3.1 Data Preparation

The Oxford Parkinson's Telemonitoring dataset was preprocessed to ensure data quality and integrity. This involved handling missing values, normalizing the data, and selecting relevant features for analysis. The dataset consists of 5,875 voice recordings from 42 individuals with early-stage PD, with each recording containing various voice features alongside motor and total UPDRS scores.

3.2 Statistical Analysis

To explore the relationship between voice measures and UPDRS scores, we utilized Ordinary Least Squares (OLS) regression analysis. The dependent variable was the motor UPDRS score, while the independent variables included demographic factors (age and sex) and various voice measures such as Jitter, Shimmer, HNR, RPDE, DFA, and PPE.

The regression model was formulated as follows:

$$\begin{aligned} \text{motor\_UPDRS} = \beta_0 + & \quad (1) \\ \beta_1 \cdot \text{Jitter}(\%) + & \quad (2) \\ \beta_2 \cdot \text{Shimmer} + & \quad (3) \\ \beta_3 \cdot \text{HNR} + & \quad (4) \\ \beta_4 \cdot \text{RPDE} + & \quad (5) \\ \beta_5 \cdot \text{DFA} + & \quad (6) \\ \beta_6 \cdot \text{PPE} + & \quad (7) \\ \beta_7 \cdot \text{age} + & \quad (8) \\ \beta_8 \cdot \text{sex} + \epsilon & \quad (9) \end{aligned}$$

where β_0 is the intercept and ϵ represents the error term. The significance of each predictor was assessed using p-values, with a threshold of $p < 0.05$ indicating statistical significance.

In addition to OLS regression, we also implemented Ridge and Lasso regression techniques to address potential multicollinearity among predictors and enhance model performance. Ridge regression applies L2 regularization by adding a penalty equal to the square of the magnitude of coefficients to the loss function. This helps to shrink the coefficients of correlated predictors towards each other.

Lasso regression, on the other hand, uses L1 regularization which adds a penalty equal to the absolute value of the magnitude of coefficients. This technique not only helps in reducing overfitting but also performs variable selection by driving some coefficients to zero, effectively excluding them from the model.

3.3 Principal Component Analysis

To further understand the underlying structure of the voice measures, we performed Principal Component Analysis (PCA). This technique was employed to reduce dimensionality while retaining as much variance as possible in the dataset. The explained variance ratio for each principal component was calculated to identify key features contributing to the variance in voice measures.

3.4 Model Evaluation

The performance of the regression models (OLS, Ridge, and Lasso) was evaluated using R-squared values to determine how well each model explained the variance in motor UPDRS scores. Additionally, diagnostic tests such as the Shapiro-Wilk test for normality and Durbin-Watson statistic for autocorrelation were conducted to validate model assumptions.

This methodology allows us to rigorously assess the predictive power of voice measures in relation to PD progression, contributing valuable insights into non-invasive monitoring techniques.

4 Results

The analysis of the Oxford Parkinson's Telemonitoring dataset revealed significant relationships between various biomedical voice measures and the motor and total UPDRS scores. The dataset consisted of 5,875 voice recordings from 42 individuals with early-stage Parkinson's disease, collected over a six-month period.

4.1 Descriptive Statistics

Table 3 presents the descriptive statistics for key variables in the dataset. The mean motor UPDRS score was 21.30 (SD = 8.13), while the mean total UPDRS score was 29.02 (SD = 10.70). The demographic characteristics of the participants included a mean age of 64.80 years (SD = 8.82) with a sex distribution of 68.2% male and 31.8% female.

Variable	Mean	Standard Deviation
Motor UPDRS	21.30	8.13
Total UPDRS	29.02	10.70
Age (years)	64.80	8.82

Table 3: Descriptive Statistics of Key Variables

4.2 Ordinary Least Squares Regression Results

The Ordinary Least Squares regression analysis yielded an R^2 value of 0.150, indicating that approximately 15% of the variance in motor UPDRS scores could be explained by the model. The results are summarized in Table 4. Significant predictors included Jitter(%), Jitter(Abs), HNR, DFA, PPE, age, and sex.

Predictor	Coefficient	Std Error	p-value
Intercept	29.5640	2.724	¡0.001
Jitter(%)	440.0807	184.349	0.017
Jitter(Abs)	-71820.00	8509.395	¡0.001
HNR	-0.4294	0.058	¡0.001
DFA	-22.6633	1.980	¡0.001
PPE	20.2435	2.547	¡0.001
Age	0.2079	0.013	¡0.001
Sex (Male)	-1.0425	0.275	¡0.001

Table 4: OLS Regression Results for Motor UPDRS Scores

The coefficients indicate that an increase in Jitter(%) is associated with higher motor UPDRS scores, while higher values of Jitter(Abs) and HNR are associated with lower scores, suggesting that increased vocal instability correlates with greater motor impairment.

4.3 Ridge and Lasso Regression Results

To further assess the predictive power of the voice measures and mitigate multicollinearity, Ridge and Lasso regression techniques were employed as well.

- **Ridge Regression**: The Ridge model demonstrated improved predictive performance with an R^2 value of approximately $R^2_{ridge} = 0.155$. - **Lasso Regression**: The Lasso model achieved an R^2 value of $R^2_{lasso} = 0.152$, effectively reducing some coefficients to zero, indicating variable selection.

The results from these models suggest that while OLS provides a reasonable fit, Ridge regression offers slightly better predictive accuracy by addressing multicollinearity among predictors.

4.4 Principal Component Analysis

Principal Component Analysis (PCA) revealed that the first principal component explained approximately 66.2% of the variance in voice measures, indicating that a significant portion of the variability can be captured by a reduced set of features.

Overall, these results highlight the potential of using voice measures as non-invasive biomarkers for monitoring Parkinson's disease progression, providing insights into the relationship between vocal characteristics and clinical indicators.

5 Discussion

The results of this study provide significant insights into the relationship between biomedical voice measures and the progression of Parkinson's disease (PD), as indicated by motor and total UPDRS scores. By analyzing the Oxford Parkinson's Telemonitoring dataset, we found that various vocal characteristics correlate with clinical indicators of disease severity, reinforcing the potential of voice analysis as a non-invasive monitoring tool for PD.

5.1 Correlation Between Voice Measures and UPDRS Scores

Our Ordinary Least Squares (OLS) regression analysis revealed that several voice features, including Jitter(%), Jitter(Abs), HNR, DFA, and PPE, were significant predictors of motor UPDRS scores. Specifically, increases in Jitter(%) were associated with higher motor UPDRS scores, indicating that greater vocal instability correlates with more severe motor impairments. This finding aligns with previous research identifying vocal characteristics as potential biomarkers for PD progression [**?**, **?**].

The coefficients obtained from the regression model suggest that Jitter(Abs) and HNR also play crucial roles in predicting motor impairment. The negative coefficient for HNR indicates that lower harmonic-to-noise ratios are associated with increased severity of motor symptoms. This finding is consistent with literature indicating that voice quality deteriorates as PD progresses, reflecting underlying neurological changes [**?**]. The relationship between these vocal features and clinical outcomes can be understood through existing theories on speech production, which highlight how neurological degeneration impacts both motor control and vocalization.

5.2 Ridge and Lasso Regression Techniques

To further assess the predictive power of the voice measures and mitigate multicollinearity among predictors, we implemented Ridge and Lasso regression techniques. The Ridge model demonstrated a slight improvement in predictive performance compared to OLS, achieving an R^2 value of approximately 0.155. Lasso regression effectively reduced some coefficients to zero, indicating variable selection and reinforcing the importance of specific voice measures in predicting PD progression.

These results underscore the utility of regularization techniques in enhancing model interpretability and performance. By identifying key predictors while controlling for multicollinearity, we can better understand the relationship between vocal features and clinical outcomes. The use of Lasso regression also highlights the potential for developing a more streamlined model focused on the most relevant voice measures.

5.3 Principal Component Analysis Insights

The Principal Component Analysis (PCA) indicated that the first principal component explained approximately 66.2% of the variance in voice measures. This suggests that a significant portion of variability in vocal characteristics can be captured by a reduced set of features. Identifying these principal components can facilitate developing more efficient monitoring tools for PD progression.

The high explanatory power of the first principal component suggests that certain underlying factors may be driving changes in vocal characteristics among individuals with PD. Future studies could explore these components further to identify specific vocal features that are most indicative of disease progression.

5.4 Limitations and Considerations

While our study provides valuable insights, it is important to acknowledge certain limitations. The dataset is specific to individuals with early-stage PD and has a relatively small sample size, which may limit the generalizability of our findings. Additionally, the lack of information on potential confounding variables—such as medications or environmental factors—could impact the accuracy of our regression models. For instance, variations in medication regimens can significantly influence both motor symptoms and voice characteristics.

Moreover, while our analysis focuses on voice measures collected through telemonitoring, it is essential to consider how different recording environments or equipment might affect voice quality. Standardizing recording conditions could help improve data consistency across studies.

5.5 Implications for Clinical Practice

The ability to monitor PD progression through non-invasive voice analysis presents a promising avenue for improving patient care. Integrating voice measures into routine clinical assessments could enable earlier detection of disease progression and more timely interventions. Moreover, telemonitoring systems utilizing voice analysis could facilitate remote patient monitoring, reducing the need for frequent in-person visits and enhancing patient convenience.

This study's findings support the hypothesis that voice measures can serve as valuable biomarkers for monitoring Parkinson's disease progression. By leveraging advanced statistical techniques and machine learning models, we can enhance our understanding of how vocal characteristics relate to clinical indicators of disease severity.

5.6 Future Directions

Future research should focus on longitudinal studies tracking changes in vocal characteristics over time concerning clinical outcomes. Additionally, exploring machine learning algorithms beyond traditional regression methods may yield even more accurate predictive models for PD progression based on voice data. Investigating additional acoustic features or integrating multimodal data sources could further enhance our understanding of how voice relates to motor function in individuals with Parkinson's disease.

Furthermore, advancements in technology—such as mobile health applications—could allow for real-time monitoring using voice data collected via smartphones or wearable devices. This would not only improve accessibility but also empower patients by providing them with tools to actively participate in their own health management.

In conclusion, our findings support using voice analysis as a non-invasive method for monitoring Parkinson's disease progression. By identifying key vocal biomarkers associated with clinical indicators, we can pave the way for improved patient management strategies that prioritize early intervention and personalized care.

References

[1] Arora, S., Baig, F., Lo, C., Barber, T. R., Lawton, M. A., Zhan, A., ... & Hu, M. T. (2022). Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. *Neurology*, 98(4), e428-e439.

[2] Frid, A., Safra, E. J., Hazan, H., Lokey, L. L., Hilu, D., Manevitz, L., ... & Sapir, S. (2021). Computational diagnosis of Parkinson's disease directly from natural speech using machine learning techniques. *IEEE Access*, 9, 20976-20988.

[3] Mei, J., Desrosiers, C., & Frasnelli, J. (2022). Machine learning classification of Parkinson's disease based on voice recordings. *Frontiers in Neurology*, 13, 831428.

[4] Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., & Moroz, I. M. (2009). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Nature Precedings*, 1-1.

[5] Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE transactions on Biomedical Engineering*, 57(4), 884-893.

[6] Tsanas, A. & Little, M. (2009). Parkinsons Telemonitoring [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5ZS3N