Helen Wright*, Michiel Postema, and Vered Aharonson

Towards a voice-based severity scale for Parkinson’s disease monitoring

https://doi.org/10.1515/cdbme-2024-2168

Abstract: The unified Parkinson’s disease rating scale, used to monitor the disease progression, is based on visual assessments of motor symptoms. Vocal manifestations of Parkinson’s disease differ from the motor ones, specifically in their rate of change with disease severity. As such, a different scale is needed to provide the voice measures of the disease severity. This study employed a dataset of voice-quality features from repeated recordings of Parkinson’s disease patients. The changes of all voice features across the categories were evaluated using one-way analysis-of-variance and support vector regression. Significant changes and marked non-linearly increasing or decreasing trends were shown for all features, for the three-categories scale. Significant changes and trends were obtained in the 12-categories scale, but only for the mild category and the severe category range of scores. The findings imply a potential for voice-based monitoring for the early and late severity stages of Parkinson’s disease that could be continuously used by patients and provide timely warnings of deterioration.

Keywords: Vocal features, disease monitoring, regression, UPDRS.

1 Introduction

Currently available treatments for Parkinson’s disease cannot cure the disease, but can alleviate symptoms and improve the patient’s quality of life. The efficacy of all treatments, however, might be improved by timely detection of a change in the disease severity [1]. A growing body of evidence highlights discernible alterations in the vocal patterns of PD patients [2]. A method able to identify these changes would be beneficial to both patients and clinicians, allowing small changes to be assessed conveniently and noninvasively, and guiding personalised treatments and interventions. To achieve this, a quantitative description of the perceived changes in PD speech patterns through explainable low-level vocal features must be developed. This work lays the foundation for a voice-based Parkinson’s disease severity scale by mapping selected vocal feature values against unified Parkinson’s disease rating scale (UPDRS) scores and assessing what relationships may be identified. In this way, the voice changes that occur as the disease worsens can be seen, tracked and quantified.

2 Materials and Methods

The data used for this study were taken from the UCI Parkinson’s Disease Telemonitoring Dataset [3]. The complete dataset comprises features extracted from 5923 sustained vowel phonations recorded from 42 PD patients at weekly intervals over a period of 6 months. The chosen features include voice quality measures which are traditionally used in speech therapy for PD [4]. Additional signal features were calculated to provide further insight into the complexity of the recorded signals [5]. A correlation analysis revealed high correlations between a number of features (correlation coefficients > 0.7) and a subset of features was used in the current analysis: The harmonics-to-noise ratio (HNR) is a measure of the amount of non-periodic noise in a recorded signal. The pitch period entropy (PPE) is calculated to quantify how well a stable pitch can be maintained.The recurrent period density entropy (RPDE) is an indicator of the voice signal’s deviation from exact periodicity. The jitter to pitch period quotient 5 (PPQ5) is a measure of the cycle-to-cycle variability of fundamental frequency of the voice cycle. The shimmer to amplitude perturbation quotient 3 (APQ3) is a measure of the cycle-to-cycle variability of the amplitude of the voice signal.

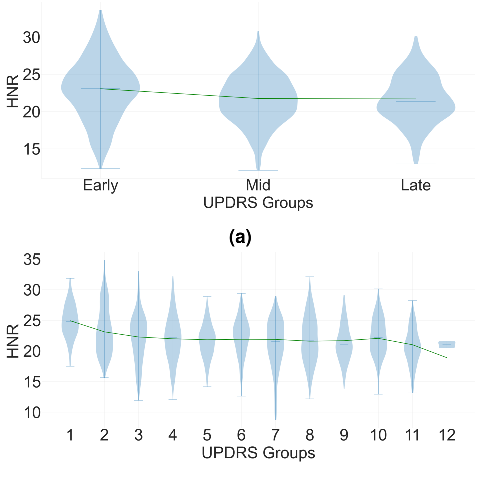
The UCI dataset includes UPDRS values, which were determined through patient examination at the beginning, midpoint and end of the data collection and recording process. To estimate missing weekly UPDRS values, linear interpolation was used. This association of UPDRS values to each recording is a unique feature of this datatset, making it particularly suitable for the current preliminary longitudinal anal-

***Corresponding author: Helen Wright,** School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, 1 Jan Smuts Laan, 2001 Braamfontein, South Africa.
e-mail: helen.wright@wits.ac.za
Michiel Postema, Department of Biomedical Technology, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland and School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, Braamfontein, South Africa.
Vered Aharonson, School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa and University of Nicosia Medical School, Nicosia, Cyprus.

ysis. Our study makes use of the motor UPDRS scores only, taken from Part III of the UPDRS assessment, since this part of UPDRS assessment includes speech-related questions. *This UPDRS score range will be referred to as UPDRS(III) in the rest of this paper*.

Two experiments were conducted to investigate how speech feature values vary with increasing UPDRS(III) scores. The range of UPDRS(III) scores in the dataset was from 5 to 40. In Experiment 1, this range was divided into three categories: early stage (scores from 5 - 15), middle stage (scores from 16 - 32), and late stage (scores from 33 - 41). These categories were first proposed in [6, 7] and are based on perceptual evaluation of PD speech recordings. In Experiment 2, the UPDRS(III) range was divided into 12 categories, with each category representing a three-point interval. This interval corresponds to the minimum number of data points required to create distinct categories that can still be linked to the ones used in Experiment 1, considering the available range of measurements in the dataset. Categories 1 to 4 (scores 5 - 16) in Experiment 2 correspond to the Early stage in Experiment 1. Categories 5 to 9 (scores 17 - 31) correspond to the Mild stage in Experiment 1. Categories 10 to 12 (scores 32 - 41) correspond to the Late stage in Experiment 1. While this division of the UPDRS(III) range did not result in perfect UPDRS alignment between the two sets of experimental groups, with a score of 16 falling into the "Early" category instead of the "Middle" category, and similarly for a score of 32, this discrepancy was allowed for this analysis as it ensured the integrity of the three-point intervals.

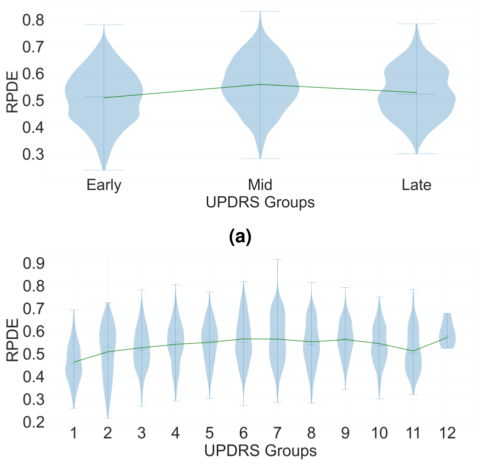
Box-and-whisker plots were drawn for each category. These were used to define the minimum and maximum values for the distributions. Any datapoints lying outside those limits were identified as outliers and removed from further analysis. The data distributions were portrayed using violin plots and their mean, median, and standard deviation were calculated. The latter descriptive statistics were compared across different UPDRS(III) categories using one-way analysis-of-variance with Bonferroni correction, for all features. Statistically significant changes, defined by a p-value of less than 0.05, were sought in each feature between the different UPDRS(III) categories. Regression analysis, using support vector regression, was conducted in both experiments to identify trends in the distribution of each feature value with increasing UPDRS(III) scores.



(a)

(b)

Fig. 1: HNR feature value distributions and regression curves. (a) shows the distribution of HNR values across three categories of UPDRS(III) scores. (b) shows the distribution across 12 UPDRS(III) categories. In both cases, SVR curves show decreasing trends.



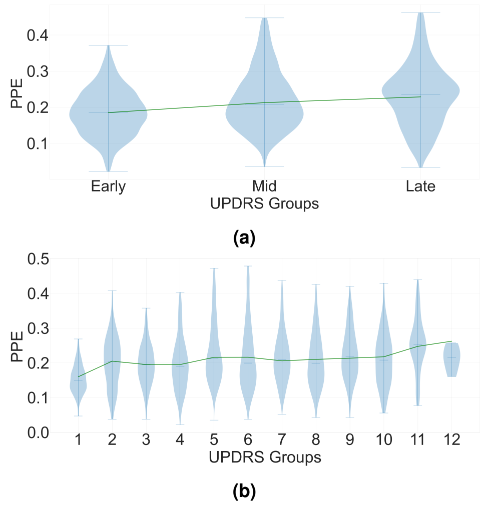
(a)

(b)

Fig. 2: RPDE feature value distributions and regression curves. (a) shows the distribution of RPDE values across three categories of UPDRS(III) scores. (b) shows the distribution across 12 UPDRS(III) categories. SVR curve shows trends that first increase, then decrease.

3 Results

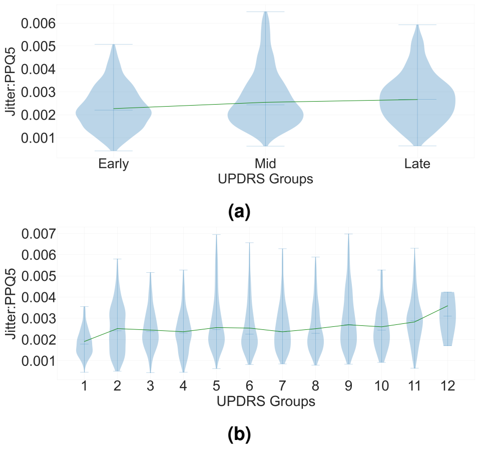
In Experiment 1, all features showed statistically significant differences between the low and middle categories, evidenced by p-values < 0.01. Three out of the five features showed statistically significant differences between the middle and high categories, namely the shimmer, PPE and RPDE - the jitter and HNR features did not. Regression curves were drawn between the categories to identify trends in feature values with increasing UPDRS(III) score. The harmonics-to-noise ratio (Figure 1a) showed a decreasing trend. Pitch period entropy (Figure 3a), jitter (Figure 4a) and shimmer (Figure 5a) all showed increasing trends. The recurrent period density entropy showed a trend that increases from the low to mid category and then decreases from the mid to high categories (Figure 2a).



(a)

(b)

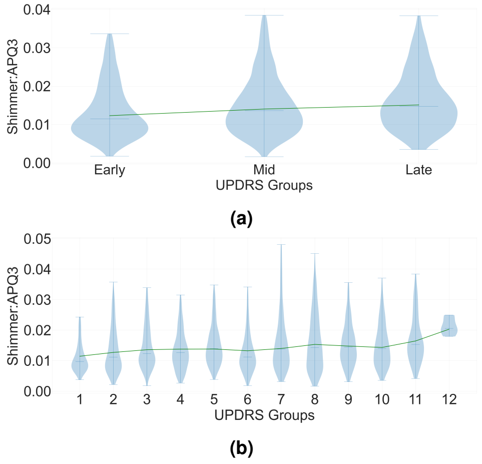
Fig. 3: PPE feature value distributions and regression curves. (a) shows the distribution of PPE values across three categories of UPDRS(III) scores. (b) shows the distribution across 12 UPDRS(III) categories. In both cases, SVR curves show increasing trends.



(a)

(b)

Fig. 4: Jitter:PPQ5 feature value distributions and regression curves. (a) shows the distribution of jitter values across three categories of UPDRS(III) scores. (b) shows the distribution across 12 UPDRS(III) categories. In both cases, SVR curves show increasing trends.



(a)

(b)

Fig. 5: Shimmer:APQ3 feature value distributions and regression curves. (a) shows the distribution shimmer values across three categories of UPDRS(III) scores. (b) shows the distribution across 12 UPDRS(III) categories. In both cases, SVR curves show increasing trends.

The use of a smaller analysis window in Experiment 2 allowed feature changes to be analysed over a finer resolution of 12 UPDRS(III) categories. The p-values of the changes in feature values between the twelve categories are given in Table 1. None of the features analysed showed significant changes between all adjacent categories. Instead, they exhibited a fluctuating pattern with increasing and decreasing values. The harmonics-to-noise ratio, pitch period entropy and recurrent period density entropy showed statistically significant differences between the first two and last three categories. The jitter and shimmer features showed minor fluctuations in the middle categories. Regression curves plotted for all features showed the same trends as those from experiment 1. However, they capture the small fluctuations between categories. The harmonics-to-noise ratio shows a gradually decreasing trend (Figure 1b). The pitch period entropy (Figure 3b), jitter (Figure 4b) and shimmer (Figure 5b) show gradually increasing trends. The recurrent period density entropy shows a trend of an increase in the early categories and then decreases in the later categories, with minor fluctuations in the middle categories (Figure 2b).

4 Discussion and Conclusions

When the UPDRS(III) range was divided into three categories, of low middle and high severity, each category showed a wide range of values, with overlap between categories. This is true for all the features analysed. However, the presence of statistically significant differences between these distributions suggests that these changes are measurable and, if observed over time, may provide a method of monitoring long-term deterioration. The large analysis windows used in this experiment, however, may obscure slighter changes within each category. The use of a smaller analysis window, as in Experiment 2, al-

Tab. 1: Statistical p-values calculated for experiment 2

Feature Name	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10	10 - 11	11 - 12
HNR	< 0.01	0.018	1.0	0.82	0.025	0.00048	1.0	0.073	0.0013	<0.01	1.0
PPE	< 0.01	1.0	1.0	<0.01	0.11	1.0	1.0	0.0034	0.84	<0.01	1.0
RPDE	< 0.01	1.0	0.00013	1.0	1.0	1.0	0.00011	0.0013	0.0091	0.094	1.0
Jitter:PPQ5	< 0.01	1.0	1.0	<0.01	0.11	1.0	1.0	<0.01	0.03	0.0045	1.0
Shimmer:APQ3	< 0.016	0.19	1.0	0.001	<0.01	<0.01	1.0	1.0	1.0	0.17	1.0

lows the changes to be observed in a finer resolution. The results highlight and confirm the gradual nature of the changes to vocal features, and provide evidence that smaller UPDRS(III) analysis windows are better for detailed analysis.

The combined findings of the 2 experiments show a tradeoff which must be considered when implementing these voice features as biomarkers of changes in PD. The low, middle and high stages are easy to interpret clinically and show consistent differences between stages but are too broad and cannot show small changes due to treatment, for example. Increasing the number of stages to 12 exhibits smaller severity changes, but only for the low and high severity stages, while the middle severity stages showed no change or fluctuative changes.

Regression analysis conducted in both experiments showed the changes in each feature with increasing UPDRS(III) score. The support vector regression techniques used for this are effective for handling non-linear relationships and are able to capture local trends. In this way, the feature value changes are captured. They also capture the gradient of the trends, which identifies the relationship between the features and disease severity. The non-linear nature of the trends further highlights the complexity of the relationship between PD progression and speech. These findings align with previous works which compared feature values between PD patients and healthy control subjects, and extends it by plotting the longitudinal changes. They also support the use of speech in the early- and late-stage assessment of PD severity.

While this work supports the use of voice for the monitoring of PD progression, it is as yet unverified. Repeating the analysis on additional datasets would serve to verify the findings and would also confirm the feature value ranges observed here. Alternative UPDRS(III) groupings could also be investigated. This would allow the optimal grouping to be identified, especially for the middle severity ranges, and allow for better alignment between the different experiment groups. This work has focused on the phonatory aspects of speech, extracted from sustained vowel phonations. Including other vocal features, extracted from alternative speaking tasks, would allow the changes in those features to be identified. These could provide additional insight into the reported vocal changes and may also be used as biomarkers.

A limitation of this study is the size of the dataset and a lack of healthy control data for comparison. Extending the analysis to a larger dataset would verify the findings reported here and allow the feature values and changes to be better quantified. A comparison with healthy control data would indicate the vocal deterioration and confirm the differences reported.

Author Statement

Research funding: This work was not supported financially. Conflict of interest: Authors state no conflict of interest. Ethical approval: The research related to human data use complied with all the relevant national regulations and institutional policies and has been approved by the University of the Witwatersrand human research ethics committee (clearance certificate numbers M221107).

References

[1] Armstrong MJ, Okun MS. Diagnosis and treatment of Parkinson disease: a review. JAMA 2020;323:548–60.

[2] Ma A, Lau KK, Thyagarajan D. Voice changes in Parkinson’s disease: what are they telling us? J Clin Neurosci 2020;72:1–7.

[3] Tsanas A, Little M. Parkinson’s Disease Telemonitoring Dataset. UCI Machine Learning Repository, 2009. Available from https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring

[4] Raphael LJ, Borden GJ, Harris KS. Speech Science Primer: Physiology, Acoustics, and Perception of Speech. Philadelphia: Lippincott Williams & Wilkins; 2007.

[5] Tsanas A, Little M,McSharry P, Ramig L. Accurate telemonitoring of Parkinson’s disease progression by non-invasive speech tests. IEEE Trans Biomed Eng 2010;57:884–93.

[6] Martínez-Martín P, Rodríguez-Blázquez C, Alvarez M, Arakaki T, Arillo VC, Chaná P, et al. Parkinson’s disease severity levels and MDS-Unified Parkinson’s Disease Rating Scale. Parkinsonism Relat Disord 2015;21(1):50–4.

[7] Sakar BE, Serbes G, Sakar CO. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson’s disease. PLoS One 2017;12:e0182428.