RESEARCH ARTICLE

Speech-based identification of L-DOPA ON/OFF state in Parkinson's Disease subjects

R. Norel PhD[1] | C. Agurto PhD[1] | J.J. Rice PhD[1†] |
B.K. Ho MD[2] | G.A. Cecchi PhD[1]

[1] IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598, USA

[2] Department of Neurology, Tufts University School of Medicine and Tufts Medical Center, 800 Washington St, Boston, MA 02111 USA

Correspondence
R. Norel PhD, IBM Computational Biology Center,P.O.Box 218, Yorktown Heights, NY 10598
Email: rnorel@us.ibm.com

Funding information
This is internally funded work from IBM and Pfizer.

Background: Parkinson's disease patients (PDP) are evaluated using the unified Parkinson's disease rating scale (UPDRS) to follow the longitudinal course of the disease. UPDRS evaluation is performed by a neurologist, and hence its use is limited in the evaluation of short-term (daily) fluctuations. Subjects taking L-DOPA as part of treatment to reduce symptoms exhibit motor fluctuations as a common complication. **Objectives:** The aim of the study is to assess the use of speech analysis as a proxy to continuously monitor PDP medication state. **Methods:** We combine acoustic, prosody, and semantic features to characterize three speech tasks (picture description, reverse counting and diadochokinetic rate) of 25 PDP evaluated under different medication states: "ON" and "OFF" L-DOPA. **Results:** Classification of medication states using features extracted from audio recordings results in cross-validated accuracy rates of 0.88, 0.84 and 0.71 for the picture description, reverse counting and diadochokinetic rate tasks, respectively. When adding feature selection and semantic features, the accuracy rates increase to 1.00, 0.96 and 0.83 respectively; thus reaching very high classification accuracy on 3 different tasks. **Conclusions:** We show that speech-based features are highly predictive

Abbreviations: MFCC, Mel-Frequency Cepstral Coefficients; PD, Parkinson's Disease. PDP, Parkinson's Disease Patients.

[†] Deceased 23 February 2018

of medication state. Given that the highest performance was obtained with a very naturalistic task (picture description), our results suggest the feasibility of accurate, non-burdensome and high-frequency monitoring of medication effects.

KEYWORDS
Parkinson's, Speech Analysis, Medication State, ON/OFF state, L-DOPA

Parkinson's Disease (PD) is the second most common neurodegenerative disease, after Alzheimer's. The estimated prevalence of PD in industrialized countries is 0.3% in the general population, 1.0% in people older than 60 years, and 3.0% in people over 80 years old [1]. Around 7 to 10 million people worldwide live with PD; that is more than the combined number of people diagnosed with multiple sclerosis, muscular dystrophy and Lou Gehrig's disease. In the US, nearly 60,000 Americans are diagnosed with Parkinson's disease each year, not including the potentially thousands of undetected cases [2]. To follow the progression of PD, the most widely used clinical rating scale is the Unified Parkinson Disease Rating Scale (UPDRS) [3]. It was originally developed in the 1980s and revised in 2001. The UPDRS was not meant for continuous monitoring; it has to be administered by neurologists or motor disorder specialists. As an alternative explored in recent years, speech can potentially be used to monitor patients, to inform on medication effectiveness, and to follow progression. The UPDRS already includes a section for scoring speech in 5 different levels: 0: Normal (no problems); 1: Slight (speech is soft, slurred or uneven); 2: Mild (occasionally parts of the speech are unintelligible); 3: Moderate (frequently parts of the speech are unintelligible); and 4: Severe (speech cannot be understood). However speech scores are inconsistent among graders according to Martinez-Martin *et al.* [4]. Therefore, there is a need for an unbiased measurement of speech changes in the research and clinical communities. Currently the most common treatment of PD includes the use of L-DOPA, which helps ameliorate the symptoms. Unfortunately long-term use of the drug results in symptom control fading out, resulting in fluctuation of medication states known as "ON" and "OFF" states [5, 6].

Speech disorder resulting from neurological impairment such as PD is known as dysarthria. This condition affects mainly the control and execution of movements related to speech production [7]. Previous studies have characterized speech in PD as having the following attributes: reduced loudness, monopitch, monoloudness, reduced stress, breathy and hoarse voice quality, and imprecise articulation [8]. Most recent approaches demonstrate that speech features can help differentiate healthy controls from PDP with high accuracy, in particular using vocal measurements of sustained phonations [9, 10, 11, 12, 13]. There is also evidence of cognitive impairment as part of PD, which affects - or is reflected in - language [14]. Yet another potential difference between PDP and healthy controls or PDP in ON/OFF state is the use of action verbs [15, 16, 17, 18]. Garcia *et al.* [19, 20] and more recently Cotelli *et al*, [21] compared PDP with healthy controls reporting action verb deficit in PDP. Herrera and Cuetos [18] added evidence that PDP without adequate dopamine levels have difficulties in naming action verbs from pictures.

However, there has not been enough research nor strong findings on the evaluation of the use of speech as a way to monitor the medication states [22, 23, 24]. Among those articles, we found that Okada *et al.* [25], analyzed isolated vowel articulation in PD subjects reporting that vowel space area was significantly expanded after L-DOPA treatment contrary to previous findings by [26], where no changes in speech over L-DOPA cycle were found. A recent article [27] analyzes a small cohort of late stage PD subjects using data from a L-DOPA challenge described in [28]. This work, which was limited to the analysis of sustained vowel /a/ and the repetition of a 8-word simple sentence, did not find significant changes in speech.

Therefore, there is a need for objective ways of evaluating PD patients as a way of monitor their disease progression, as well as the effect and duration of their medication. Perhaps there is a prodromal signature in speech that could inform subjects about the risk of PD so they could seek treatment to slow its progression.

In this paper we show that features extracted from speech are informative enough to distinguish the medication state of a PD subject. In particular, we show that a simple and naturalistic task, namely the description of a picture, provides for the highest accuracy, suggesting a potential use in high-frequency and remote monitoring.

SUBJECTS AND METHODS

| Subjects

Twenty five subjects (6 females with age of 67 ± 6 years; 19 males with age of 69 ± 7.5 years) with idiopathic Parkinson's disease. The study was approved by the Tufts University Institutional Review Board.

Inclusion criteria: subjects that respond to L-DOPA treatment, are able to recognize their "wearing off" symptoms, can confirm that they usually improve after their next dose of PD medication, have PD Hoehn & Yahr Stage less than or equal to 3 (assessed while the patient is "ON"), and a score of 26 or more on the Montreal Cognitive Assessment Tool (MoCA) which is the normal range (i.e., no cognitive impairment). Exclusion criteria includes any current history of neurological disease (except for Parkinson's disease), cognitive impairment, or psychiatric illness that in the investigator's judgment would interfere with subject participation, treatment with an investigational drug within 30 days, or 5 half lives preceding the first dose of study treatment, whichever is longer, history of regular alcohol consumption exceeding 7 drinks/week for females or 14 drinks/week for males, subjects with cardiac pacemakers, electronic pumps or any other implanted medical devices (including deep brain stimulation devices). Each subject was evaluated by a neurologist using the UPDRS. The differences of speech scores between "OFF" and "ON" states were 2 (1 subject), 1 (7 subjects), 0 (16 subjects), and -1 (1 subject).

| Design and Protocol

This study reflects the analysis of speech tasks from "Observational Study in Parkinson's Patient Volunteers to Characterize Digital Signatures Associated with Motor Portion of the UPDRS, daily living activities and speech", conducted by IBM, Pfizer and Tufts. In this study, three different speech tasks were acquired for each subject at two different sessions: before and after their L-DOPA medication. The order of medication state across sessions was randomized to decrease the learning effect that may influence the results. The first task is called "Picture description" (cookie theft [29, 30] or description of another picture of similar characteristics). In this task, the participant is asked to provide a free-form verbal description of a visual stimulus (the picture). For each session, a different picture was presented to the subject, picture_1 for all subjects in visit_1 and picture_2 for all subjects in visit_2. The second task is a modification of a classic test for mental state evaluation [31, 32] and its called "Reverse counting". Participants were asked to count in reverse order starting from a different number in each of the sessions to keep the same level of cognitive load. The third task is called "Diadochokinetic rate" and the subjects were asked to pronounce the sequence of three syllables "pa-ta-ka" as rapidly as possible for 10 seconds. This test is widely used for assessing oral motor skills [33].

| Data Acquisition

The speech tasks were recorded using Audacity software [34] in two channels, one for the analyzed subject and one for the experimenter. The recording parameters were set to sampling frequency of 44.1 kHz with 16 bits and were saved using 'wav' format. Both the subject and the experimenter had to wear a headset with a low-impedance unidirectional dynamic microphone.

| Feature Extraction

The characterization of the speech recordings is performed using two software tools: Python [35] and Praat [36, 37]. Three types of features, which are explained below, were extracted to find signatures of different medication states in the three speech tasks.

| Mel Frequency Cepstral Coefficients (MFCCs)

Thirteen MFCCs were calculated using python_speech_features package [38]. Following common practice [39], the first coefficient was replaced by the log of the total frame energy in order to analyze the overall energy in the speech of the speaker. The parameters used to calculate the coefficients were windows of 25 ms and windows overlap of 10 ms. To only characterize the voice of the subject, pauses were removed from the recording. A pause is defined by a silence threshold of -25 dB and minimum duration of 100 ms as recommended by Griffiths [40]. To represent the distribution of each coefficient, we computed 10 statistical descriptors, mean, variance, kurtosis, skewness, mode, percentiles 10th, 25th, 50th, 75th, and 90th. These features are used for all analyzed speech tasks.

| Nuclei syllable (NS)

As a proxy for fluency in speech (or speech rate), we used the method proposed by De Jong *et al.*[41], to estimate the location of syllables. After locating the syllables, we computed the time lapse between them in the speech recording with and without pauses being removed. To represent the distribution of syllable duration, we computed the following statistical descriptors: mean, variance, kurtosis, skewness, mode, interquartile range (IQR, a measure of variability), 10th percentile, and 90th percentile for a total of 16 features. These features are calculated for the three speech tasks.

| Semantic features (SF)

Since the picture description task uses free speech, we analyzed the semantic content of the description provided by the subjects. After manually transcribing the recordings, we used the Stanford POS tagger [42] to get part of speech tags for words uttered by subjects. For nouns and verbs we computed the similarity distance to the following seed words: *action*, *act*, *move*, *play*, *energetic*, *inaction*, *sleep*, *rest*, *sit* and *wait*. The choice of words was aimed to check whether the use of action vs. non action words was influenced by dopamine level ("ON" vs. "OFF" state of L-DOPA). Words are represented using Gloval vectors for word representations (GloVe) [43]. Briefly, GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. We used GloVe version 1.2, using vector representation of 300 dimensions, trained on six billion word corpus taken from Wikipedia 2015 and Gigaword 5 [44]. The distance similarity is computed between each verb and noun uttered by the subject and each of the seed words. To represent the distribution obtained for each seed words, we calculated the following statistical descriptors from the distances of the subjects words: median, 10^{th} percentile, 90^{th} percentile, skewness, kurtosis, IQR. As an additional feature, we also compute the total number of words used in the analysis.

| Statistical Analysis

Features are ranked based on p-values obtained after performing a two sample t-test. This procedure is applied separately for the training sets generated in the validation procedure. To get an insight of how the features interact in each medication state, we also computed the partial correlations among the top features for each speech task. Partial correlation captures the pattern of covariation between a pair of features by removing the effect of the other analyzed features.

| Classification

We evaluate the potential of our features to differentiate one medication state from another by applying different classifiers. We chose 4 different classifiers: Nearest Neighbors (NN), Logistic Regression (LR) with l_1-norm regularization, SVM with elastic regularization and Random Forests (RF). To avoid gender, age and education level confounding effect, we use each subject as his/her own baseline. This means that instead of performing "ON" vs. "OFF" classification, we performed ("ON" - "OFF") vs ("OFF" - "ON") classification. Before providing the features to the classifiers, the features are standardized (mean = 0 and standard deviation = 1). Finally, accuracy rates are calculated using a two nested cross-validation approach and we provide their 95% confidence intervals (CI) after performing bootstrap with 1000 samples.

| Impact of Audio Duration

To evaluate whether a longer recording can provide more informative features, we analyze subjects with recordings of more than 20 seconds in the picture description task (after removing pauses in "ON" and "OFF" states). For each recording, we extract windows of 5, 10, 15, 18 and 20 seconds centered at half of the recording and calculate only MFCC features. To evaluate the impact of the length of the recording on the accuracy rate, we perform 4-fold cross validation based on 100 different data partitions.

RESULTS

| Statistical analysis

Table 1 shows the top 5 features for each speech task, where only acoustic features, specifically MFCC #1 (total energy) and MFCC #11, appear to be the most relevant for reverse counting and diadochokinetic rate tasks, respectively. On the other hand top-ranked features for picture description shows SF, NS and MFCC features. For this reason, we only evaluated patterns of co-variation using partial correlation among the top features for the picture description task, which are shown in Fig. 1. It can be seen that the "OFF" state is characterized by a strong positive partial correlations between SF (*play*) and acoustic (MFCC #2) and SF (*act* and *play*) and NS features.

TABLE 1 Five top-ranked features for "ON" vs "OFF" states characterization for each speech task. Ranking is calculated with all of the extracted features using two-sample t-test in each training set of our leave-subject-out cross validation approach. Listed statistics are estimated in all samples for reference only. A positive t-statistic indicates greater mean value for the ON state.

Speech Task	Feature	p-value	t-statistic
Picture description	PLAY (pct10)	5.92e-07	5.76
	MFCC #2 (md)	1.50e-05	4.82
	ACT (pct10)	2.79e-05	4.63
	MFCC #12 (sk)	3.96e-04	-3.81
	NS (pct90)	5.10e-04	3.73
Reverse counting	MFCC #1 (q50)	1.54e-07	-6.14
	MFCC #1 (q25)	9.04e-07	-5.63
	MFCC #1 (mn)	4.09e-06	-5.20
	MFCC #1 (q75)	4.88e-06	-5.15
	MFCC #8 (sk)	3.11e-05	4.60
Diadochokinetic rate	MFCC #11 (pct75)	2.46e-05	4.69
	MFCC #11 (mn)	1.06e-04	4.24
	MFCC #11 (pct50)	1.86e-04	4.06
	MFCC #3 (pct75)	2.57e-03	-3.19
	MFCC #11 (pct25)	2.74e-03	3.17

| Classification

As explained above, classification tasks are performed by subtracting features from one state to another state. Table 2 shows the classification accuracy using all features and feature selection for the different possible combinations of features in each speech task. LR with l_1-norm regularization is the classifier that helps to improved performance for several feature combinations. Fig. 2 shows the best performance among the combinations of features (highlighted in Table 2). Results using ten-fold cross-validation are indicated in the barplots. The CI is shown, and all achieve results higher than chance probability.

| Impact of Audio Duration

Figure 3 shows the effects in accuracy of fixing the duration of the speech recording before feature extraction. Median accuracy value of 0.70 using 16 subjects in comparison of 0.88 using 25 subjects is achieved for recording length of 10 seconds or more. For this experiment we use exclusively the acoustic features and only one of the possible combinations of classifier and number of selected features, chosen based on the results presented on Table 2 (LR - l_1-norm using 130 features).

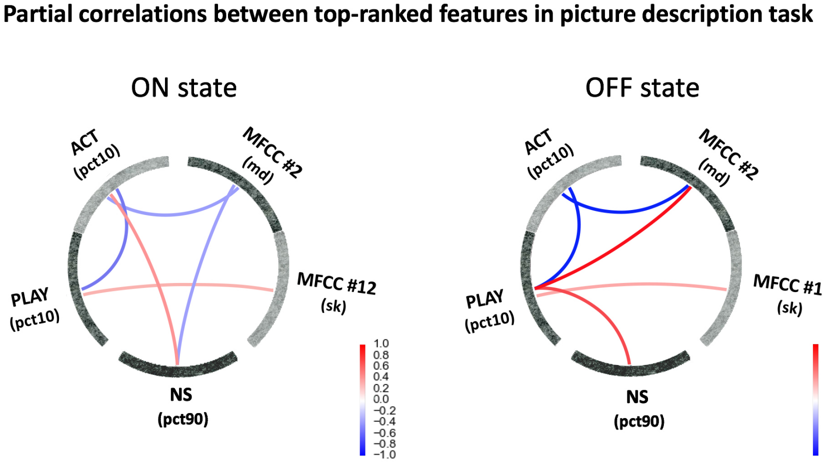

FIGURE 1 Partial correlations for "ON" and "OFF" states were calculated using the top 5 features of the picture description task described in Table 1. Positive correlations are displayed in red color while negative correlations are in blue. "OFF" state shows a stronger correlation among features in comparison with "ON" state. In addition, new positive correlations were found between SF features (play), NS and MFCC #2

DISCUSSION

High accuracy rates with values above chance (see Fig. 1) were achieved for all speech tasks, in particular for picture description (1.0) and reverse counting (0.96). This is consistent with the work in [45], which suggests that information extracted from running speech is better to detect PD signatures than information acquired using a diadochokinetic rate task. Furthermore, Ackermann *et al.* [33] reported that for PD subjects there may be a trade-off between amplitude of articulator movement and rate of speech affecting the results of diadochokinesis tests.

The most predominant features for the three speech tasks were MFCCs. These features can make a better characterization of the voice as they analyze it using different frequency bands. We observe in Fig. 1 that MFCCs #1 and #11 are the most significant features for reverse counting and diadochokinetic rate tasks, respectively. MFCC #1 captures information from the total frame energy, meaning that in this task the main difference between medication states is characterized by changes in the speech energy. This speech energy variation is one of the characteristics of hypokinetic dysarthria found in PDP and reported in [46]. On the other hand MFCC #11 captures information in high frequency [9.5kHz - 12.6kHz]. Recently researchers, supported by the advancements in equipment technology that captures a broader spectrum [47, 48], have shown that there is perceptually relevant information on high frequency speech, affecting speech intelligibility. Both [47, 49] concluded that high frequency speech characteristics are different in dysphonic versus control subjects, suggesting that the hoarseness characteristic of PD subjects, or in our case, the difference in hoarseness between "ON" and "OFF" states is what we are capturing with high coefficients of the MFCC. There is a report on Parkinson's rat models [50] that finds a similar trend of what is seen on Table 1. In the rat model, the control rats had a higher maximum frequency than the dopamine-altered (reduced) rats for both the simple and frequency modulated calls.

In the picture description, the three type of features are informative and complement each other. By studying the covariation pattern shown in Fig. 1, we observe that a very high positive correlation (red lines) occurs only in "OFF" state between SF (*play*) and the other to types of features MFCC #2 and NS. This interaction observed in only one

TABLE 2 Performance achieved in each task for different combinations of features. Accuracy rates are calculated with and without feature selection. Only the classifiers with highest accuracy value are shown. The highest accuracy rate is obtained for picture description with and without feature selection. MFCC features are relevant for achieving good performance in the different speech tasks. LR - l_1-norm achieves improved performance for several combinations possibly due to the sparse nature of the features

Speech Task (number of subjects)	Features	Accuracy using all features	Classifier for all features	Best accuracy (# features)	Classifier for best accuracy
Picture description (25 subjects)	NS	0.64 (16)	NN	0.76 (15)	NN
	SF	0.68 (61)	LR - l_1-norm	0.84 (1)	NN
	NS + SF	0.80 (77)	LR - l_1-norm	0.84 (1)	NN
	MFCC	0.88 (130)	LR - l_1-norm	0.88 (130)	LR - l_1-norm
	MFCC + NS	0.80 (146)	LR - l_1-norm	0.84 (27)	LR - l_1-norm
	MFCC + SF	0.76 (191)	LR - l_1-norm	1.0 (5)	LR - l_1-norm
	MFCC + SF + NS	0.72 (207)	LR - l_1-norm	1.0 (5)	LR - l_1-norm
Reverse counting (25 subjects)	NS	0.76 (16)	NN	0.76 (16)	NN
	MFCC	0.76 (130)	RF	0.88 (82)	NN
	MFCC + NS	0.84 (146)	SVM - elastic	0.96 (15)	LR - l_1-norm
Diadochokinetic rate (24 subjects)	NS	0.71 (16)	RF	0.71 (15)	RF
	MFCC	0.67 (130)	LR - l_1-norm	0.83 (1)	NN
	MFCC + NS	0.63 (146)	SVM - elastic	0.83 (1)	NN

state between features help to achieve 12% improvement with respect to use only MFCC features (see Table 2). It is interesting to note on Table 1 that distance to seed words *play* and *act*, which both denote *activity* are more prominent in the ON state than the OFF state (positive sign on t-statistic).

SF are informative; however they require the context from the speaker for these to be understood. Therefore, this task needs to be designed properly. On the other hand, acoustic features are more flexible for the analysis as any part of the speech recording can be used to characterize the voice.

In addition, we also perform an extra experiment with the picture description task data to evaluate the impact of duration of the analyzed recording in the accuracy. The accuracy values are reduced from 0.88 to 0.70 as only 16 out of 25 subjects were used and 4-fold cross validation was implemented instead of leave-one-out. Nevertheless, we observe in Fig. 3 that the accuracy results are very stable when 10 or more seconds of recording are analyzed. This small duration makes feasible the implementation in mobile applications that can be used as part of a daily task to monitor PD subjects.

The experimental design balanced the visit order ("ON" in first visit 50% of the cases) to avoid interference with the classification results and the visit order. Nevertheless we tested classification using the visit order as the class. For syllable repetition where the task is exactly the same in both visits the accuracy rate is 0.52 (just acoustic features) and 0.50 (acoustic plus prosody features), thus we can assume that there is not much to learn from one visit to the next. For the reverse counting task, the accuracy rates are 0.68 (just acoustic features) and 0.64 acoustic plus prosody features; even though the task is not identical in both visits since the initial number is different, in both cases is a reverse counting

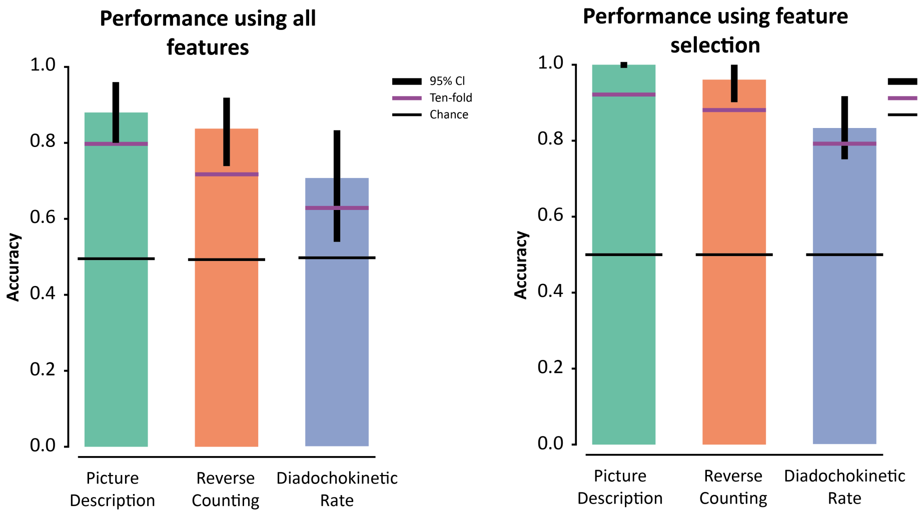

FIGURE 2 Classification performance for each task using all features (left figure) and after performing feature selection (right figure) using leave-one-subject-out cross validation. Confidence intervals at 95% are marked with black vertical lines. Chance probability is calculated for each speech task and displayed with a horizontal black line. We also indicate the accuracies obtained for ten-fold cross validation with horizontal purple lines. Results obtained with leave-one-subject-out and ten-fold cross validation surpass chance probability.

using 3 as decrement unit, we interpret the results as the second time the task should be more familiar than the first time thus some learning may occur. Better accuracy rates are obtained for the picture description task, 0.84 both when using just acoustic features and when using acoustic and prosodic features; we can speculate on two reasons for this fact, there could be some learning effect, you are describing a picture both times or the medicine state is confounded with the picture used in each visit; for example the picture used for visit one, regardless of perceived L-DOPA state (ON/OFF) is easier (harder) to describe with respect to the picture used for the second visit. The conclusion from this experiment is that our models, which are different when classifying visit order or medicine ON/OFF state, are robust enough to overcome any learning effect or confounding effect with visit order and are able to distinguish L-DOPA state differences.

To find the possible causes for misclassification, we inspected the most frequently misclassified subjects. We observed, among the top reasons, that misclassified subjects present in their recordings high level of saturation and background noise produced by variation in the settings of the preamplifier. Further work will address these technical limitations, and potentially better performance may be achieved.

Finally, we want to mention that in our dataset, there is not much variation in UPDRS speech scores between medication states even though the subjects present plenty of variation in their speech. In fact, for 64% of the subjects the difference is 0, while for one case there is an improvement in speech score in OFF state. When we compared the correlation between the UPDRS total score difference and the UPDRS speech score difference, we obtained a low value of R^2 = 0.247 with p-value of 0.01. Therefore, we want to emphasize that a new quantitative metric to monitor the patient are promising to see both disease progression and the effect on the medication(s) used. Mathemati-

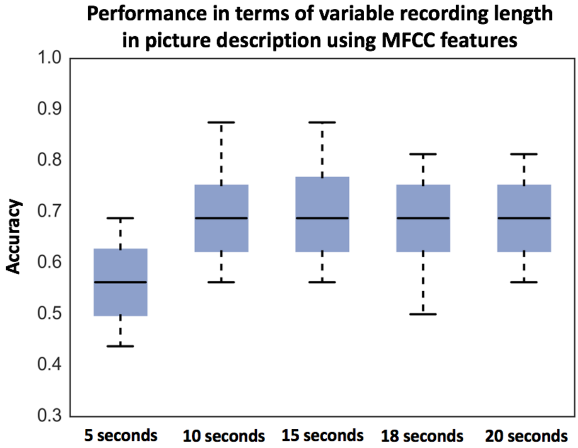

FIGURE 3 Accuracy while fixing the recording time after pause removal for 5, 10, 15, 18 and 20 seconds. Only 16 subjects were used for this analysis since we used recordings of 20 or more seconds of duration after removing pauses. The features used in this analysis were MFCC given the independence of context. Boxplots were created based on 4-fold cross validation on 100 random samples. Whiskers of boxplots show 5th and 95th percentiles. Median accuracy rate is stable around 0.70 for different recording lengths except for 5 seconds.

cal/computational analysis of speech can increase the granularity in the assessment and also avoid the human biases that results in inconsistencies between graders.

CONCLUSIONS AND LIMITATIONS

Our study explored different speech tasks that are easy to implement as a daily routine for monitoring PDP, obtaining high accuracy rates in detecting medication states. Our best results are obtained in the picture description task, which is a type of free speech. MFCC features are well known for capturing emotions [51, 52] and we believe that this fact may helped improving the classification accuracy since subjects can express emotions while describing the picture. Overall, our accuracy results range from 83% to 100% on naturalistic speech tasks demonstrate the potential of our analyses to be used as proxy to monitor subjects on a daily basis.

Given that this study involves a small cohort of PD subjects, a large study where different sites to enroll patients are involved will be required for further validation. In addition, the study will need to include higher variability in levels of speech score as well as higher differences of score between medication states.

To the best of our knowledge, this is the first paper to combine acoustic and semantic features of speech to monitor PD medication state. This opens the real possibility for continuous, unobtrusive, remote patient state monitoring.

ACKNOWLEDGEMENTS

We thank Stephen Heisig for technical help with data access and Rachel Ostrand with help with protocol and audio specs.

CONFLICT OF INTEREST

Nothing to report.

REFERENCES

[1] Lee A, Gilbert RM. Epidemiology of Parkinson disease. Neurologic clinics 2016;34(4):955–965.

[2] "Parkinson's Disease Foundation"; 2017. {http://www.pdf.org/parkinson_statistics/}, [Online; accessed 26-September-2017].

[3] Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, 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 2008;23(15):2129–2170.

[4] Martínez-Martín P, Gil-Nagel A, Gracia LM, Gómez JB, Martínez-Sarriés J, Bermejo F. Unified Parkinson's disease rating scale characteristics and structure. Movement disorders 1994;9(1):76–83.

[5] Lees A. The on-off phenomenon. Journal of Neurology, Neurosurgery & Psychiatry 1989;52(Suppl):29–37.

[6] wearingoff; 2017. {http://www.wearingoff.eu/wearing-off/describing-wearing-off}, [Online; accessed 18-October-2017].

[7] Pinto S, Ozsancak C, Tripoliti E, Thobois S, Limousin-Dowsey P, Auzou P. Treatments for dysarthria in Parkinson's disease. The Lancet Neurology 2004;3(9):547–556.

[8] Logemann JA, Fisher HB, Boshes B, Blonsky ER. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. Journal of Speech and hearing Disorders 1978;43(1):47–57.

[9] Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO, et al. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE transactions on biomedical engineering 2009;56(4):1015–1022.

[10] Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Transactions on Biomedical Engineering 2012;59(5):1264–1271.

[11] Yang S, Zheng F, Luo X, Cai S, Wu Y, Liu K, et al. Effective dysphonia detection using feature dimension reduction and kernel density estimation for patients with Parkinson's disease. PloS one 2014;9(2):e88825.

[12] Belalcázar-Bolaños EA, Orozco-Arroyave JR, Vargas-Bonilla JF, Haderlein T, Nöth E. Glottal Flow Patterns Analyses for Parkinson's Disease Detection: Acoustic and Nonlinear Approaches. In: International Conference on Text, Speech, and Dialogue Springer; 2016. p. 400–407.

[13] Garcia AM, Carrillo F, Orozco-Arroyave JR, Trujillo N, Vargas Bonilla JF, Fittipaldi S, et al. How language flows when movements don't: An automated analysis of spontaneous discourse in Parkinson's disease. Brain Lang 2016;162:19–28. Garcia, Adolfo M Carrillo, Facundo Orozco-Arroyave, Juan Rafael Trujillo, Natalia Vargas Bonilla, Jesus F Fittipaldi, Sol Adolfi, Federico Noth, Elmar Sigman, Mariano Fernandez Slezak, Diego Ibanez, Agustin Cecchi, Guillermo A ENG 2016/08/09 06:00 Brain Lang. 2016 Aug 5;162:19-28. doi: 10.1016/j.bandl.2016.07.008.

[14] Auclair-Ouellet N, Lieberman P, Monchi O. Contribution of language studies to the understanding of cognitive impairment and its progression over time in Parkinson's disease. Neuroscience & Biobehavioral Reviews 2017;.

[15] Rodríguez-Ferreiro J, Menéndez M, Ribacoba R, Cuetos F. Action naming is impaired in Parkinson disease patients. Neuropsychologia 2009;47(14):3271–3274.

[16] Fernandino L, Conant LL, Binder JR, Blindauer K, Hiner B, Spangler K, et al. Parkinson's disease disrupts both automatic and controlled processing of action verbs. Brain and language 2013;127(1):65–74.

[17] Fernandino L, Conant LL, Binder JR, Blindauer K, Hiner B, Spangler K, et al. Where is the action? Action sentence processing in Parkinson's disease. Neuropsychologia 2013;51(8):1510–1517.

[18] Herrera E, Cuetos F. Semantic disturbance for verbs in Parkinson's disease patients off medication. Journal of Neurolinguistics 2013;26(6):737–744.

[19] García AM, Ibáñez A. Words in motion: Motor-language coupling in Parkinson's disease. Translational Neuroscience 2014;5(2):152–159.

[20] García AM, Carrillo F, Orozco-Arroyave JR, Trujillo N, Bonilla JFV, Fittipaldi S, et al. How language flows when movements don't: An automated analysis of spontaneous discourse in Parkinson's disease. Brain and language 2016;162:19–28.

[21] Cotelli M, Manenti R, Brambilla M, Borroni B. The role of the motor system in action naming in patients with neurodegenerative extrapyramidal syndromes. Cortex 2017;.

[22] Jiang J, Lin E, Wang J, Hanson DG. Glottographic measures before and after levodopa treatment in Parkinson's disease. The Laryngoscope 1999;109(8):1287–1294.

[23] Goberman A, Coelho C, Robb M. Phonatory characteristics of parkinsonian speech before and after morning medication: the ON and OFF states. Journal of communication disorders 2002;35(3):217–239.

[24] Skodda S, Grönheit W, Schlegel U. Intonation and speech rate in Parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission. Journal of Voice 2011;25(4):e199–e205.

[25] Okada Y, Murata M, Toda T. Effects of Levodopa on Vowel Articulation in Patients with Parkinson's Disease. Kobe J Med Sci 2015;61(5):E144–E154.

[26] Poluha P, Teulings HL, Brookshire R. "Handwriting and speech changes across the levodopa cycle in Parkinson's disease". Acta psychologica 1998;100(1):71–84.

[27] Fabbri M, Guimarães I, Cardoso R, Coelho M, Guedes LC, Rosa MM, et al. speech and Voice response to a levodopa challenge in late-stage Parkinson's Disease. Frontiers in neurology 2017;8:432.

[28] Fabbri M, Coelho M, Abreu D, Guedes L, Rosa M, Costa N, et al. Parkinsonism and Related Disorders 2016 1;.

[29] Kaplan E. The assessment of aphasia and related disorders, vol. 2. Lippincott Williams & Wilkins; 1983.

[30] McNamara P, Obler LK, Au R, Durso R, Albert ML. Speech monitoring skills in Alzheimer's disease, Parkinson's disease, and normal aging. Brain and Language 1992;42(1):38–51.

[31] Hayman M. Two minute clinical test for measurement of intellectual impairment in psychiatric disorders. Archives of Neurology & Psychiatry 1942;47(3):454–464.

[32] Smith A. The serial sevens subtraction test. Archives of Neurology 1967;17(1):78–80.

[33] Ackermann H, Konczak J, Hertrich I. The temporal control of repetitive articulatory movements in Parkinson's disease. Brain and language 1997;56(2):312–319.

[34] Audacity;. http://www.audacityteam.org, [version 2.1.2].

[35] "Python Software Foundation";. {http://www.python.org}.

[36] Boersma P. Praat, a system for doing phonetics by computer. Glot International 2001;5(9/10):341–345.

[37] Boersma P, Weenink D, Praat: doing phonetics by computer; 2017. {http://www.praat.org/}, [Version 6.0.36, retrieved 11 November 2017].

[38] Lyons J, python-speech-features; 2016. {http://python-speech-features.readthedocs.io/en/latest}.

[39] Jurafsky D, Martin JH. Speech and Language Processing (2Nd Edition). Upper Saddle River, NJ, USA: Prentice-Hall, Inc.; 2009.

[40] Griffiths R. Pausological Research in an L2 Context: A Rationale, and Review of Selected Studies. Applied Linguistics 1991;12(4):345–364.

[41] De Jong NH, Wempe T. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods 2009;41(2):385–390.

[42] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 Association for Computational Linguistics; 2003. p. 423–430.

[43] Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1532–1543. http://www.aclweb.org/anthology/D14-1162.

[44] Pennington J, Socher R, Manning CD, GloVe.6B; 2017. [Online; accessed 11-August-2017]. https://nlp.stanford.edu/projects/glove/.

[45] Khan T, Running-speech MFCC are better markers of Parkinsonian speech deficits than vowel phonation and diadochokinetic; 2014. {http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A705196&dswid=-7062}.

[46] Gomez-Vilda P, Palacios-Alonso D, Rodellar-Biarge V, Alvarez-Marquina A, Nieto-Lluis V, Martinez-Olalla R. Parkinson's disease monitoring by biomechanical instability of phonation. Neurocomputing 2017;255:3 – 16. http://www.sciencedirect.com/science/article/pii/S0925231217305489, bioinspired Intelligence for machine learning.

[47] Monson BB, Hunter EJ, Lotto AJ, Story BH. The perceptual significance of high-frequency energy in the human voice. Frontiers in Psychology 2014;5:587. https://www.frontiersin.org/article/10.3389/fpsyg.2014.00587.

[48] Vitela AD, Monson BB, Lotto AJ. Phoneme categorization relying solely on high-frequency energy. The Journal of the Acoustical Society of America 2015;137(1):EL65–EL70.

[49] Naranjo NV, Lara EM, Rodríguez IM, García GC. High-frequency components of normal and dysphonic voices. Journal of Voice 1994;8(2):157–162.

[50] Ciucci MR, Ahrens AM, Ma ST, Kane JR, Windham EB, Woodlee MT, et al. Reduction of dopamine synaptic activity: degradation of 50-kHz ultrasonic vocalization in rats. Behavioral neuroscience 2009;123(2):328.

[51] Nwe TL, Wei FS, De Silva LC. Speech based emotion classification. In: TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology, vol. 1 IEEE; 2001. p. 297–301.

[52] Koolagudi SG, Rao KS. Emotion recognition from speech: a review. International journal of speech technology 2012;15(2):99–117.

DOCUMENTATION OF AUTHOR ROLES

Research project: Conception: JJR, GAC, RN; Data provider: BKH; Statistical Analysis: RN, CA; Review and Critique: GAC; Manuscript Preparation: RN, CA; Writing of the first draft: RN, CA; Review and Critique: GAC, JJR.