Automatically Detecting Errors and Disfluencies in Read Speech to Predict Cognitive Impairment in People with Parkinson's Disease

Amrit Romana[1], *John Bandon*[1], *Matthew Perez*[1], *Stephanie Gutierrez*[2], *Richard Richter*[2], *Angela Roberts*[2], *Emily Mower Provost*[1]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, Michigan, USA
[2]Communication Sciences and Disorders, Northwestern University, Evanston, Illinois, USA

aromana@umich.edu, jbandon@umich.edu, mkperez@umich.edu,
stephanie.gutierrez@northwestern.edu, richardrichter2020@u.northwestern.edu,
angela.roberts@northwestern.edu, emilykmp@umich.edu

Abstract

Parkinson's disease (PD) is a central nervous system disorder that causes motor impairment. Recent studies have found that people with PD also often suffer from cognitive impairment (CI). While a large body of work has shown that speech can be used to predict motor symptom severity in people with PD, much less has focused on cognitive symptom severity. Existing work has investigated if acoustic features, derived from speech, can be used to detect CI in people with PD. However, these acoustic features are general and are not targeted toward capturing CI. Speech errors and disfluencies provide additional insight into CI. In this study, we focus on read speech, which offers a controlled template from which we can detect errors and disfluencies, and we analyze how errors and disfluencies vary with CI. The novelty of this work is an automated pipeline, including transcription and error and disfluency detection, capable of predicting CI in people with PD. This will enable efficient analyses of how cognition modulates speech for people with PD, leading to scalable speech assessments of CI.

Index Terms: Parkinson's disease, cognitive impairment, read speech analysis, speech errors, disfluencies

1. Introduction

Parkinson's disease (PD) is a progressive central nervous system disorder affecting more than 2-3% of the population over 65 years of age [1, 2]. PD is primarily associated with motor symptoms, such as tremors and slow movements. More recent studies are finding that between 33-57% of people with PD also suffer from cognitive symptoms [3, 4, 5]. Understanding the effect of PD on motor versus cognitive systems is crucial for effective treatment planning, as some treatments offer a trade-off between lessening the two types of symptoms [6].

An extensive body of work has used speech to assess motor symptom severity in people with PD [7, 8, 9, 10, 11, 12, 13]. Speech is also modulated by changes in cognition, such as working memory and language planning. This points to the potential of using speech to assess cognitive symptom severity in people with PD, but only a small set of work has explored this. Thies et al. used prosodic prominence measures to predict cognitive impairment (CI) in people with PD [14], but these speech features need to be extracted by trained phoneticians, making this approach difficult to scale. For a similar prediction task, Rektorova et al. automatically extracted acoustic features, such as fundamental frequency and energy variation [15]. However, these acoustic features generally describe motor changes and are not targeted toward capturing CI.

Speech errors, such as substituting one word for another, and disfluencies, such as repeating oneself, have been demonstrated to be strong predictors of CI [16, 17]. While these errors and disfluencies are rare and difficult to capture in spontaneous speech [18], reading passage tasks can be used to assess cognition in a more controlled manner. Gollan and Goldrick recently introduced a reading passage that includes grammatical errors, prompting readers to produce errors and disfluencies [19]. Using this passage, they found that the errors readers made differed by age. For example, while older adults produced more errors in general, younger adults were more likely to revise their speech and insert a correction after making errors. However, the participants in this study did not have CI or PD. It is unknown if these error and disfluency patterns relate to CI, or, more specifically, to CI in people with PD. It is also unknown if this analysis could be automated, as relying on trained transcribers to manually annotate errors and disfluencies is both an error-prone and time-consuming process that is difficult to scale.

In this paper, we present the first analysis of how errors and disfluencies in reading change as a function of CI in people with PD. We start by analyzing errors and disfluencies in manual transcripts with annotations specifying where and which errors and disfluencies occur. Next, we present an approach for detecting the same errors and disfluencies from manual transcripts when annotations are not available. Finally, we couple this method with automatic speech recognition (ASR) to both automatically generate transcripts and detect errors and disfluencies. In each approach, we find complex error and disfluency patterns that are strong predictors of CI. This method offers insight into how cognition impacts speech for people with PD and could potentially serve as another tool for assessing CI.

2. Dataset

The data include a collection of audio recordings and transcripts from 37 people with PD who have a PD diagnosis based on the UK Brain Bank criteria [20] and at least a 10th grade education.

Participants of this study read a three-paragraph passage introduced by Gollan and Goldrick to induce speech errors and

P1	with the light of the two oil lamps the evil went away
P2	with the lamps of the two oil animals the light went away
P3	the with of light two the lamps oil evil the away went

Table 1: *The first sentence from each paragraph of the reading task. Paragraph 1 is a standard paragraph. Paragraphs 2 and 3 include swapped words to elicit reading errors and disfluencies.*

disfluencies [19]. The first paragraph resembled a standard English passage. In the second paragraph, nouns were swapped across sentences. In the third and most difficult paragraph, every consecutive pair of words was transposed. To illustrate, we list the first sentence of each paragraph in Table 1.

The readings were recorded with head-mounted microphones. These recordings were sampled at 44.1 kHz with 16-bit resolution. We downsample the audio to 16 kHz.

These readings were manually transcribed in the CHAT format by trained listeners in the Language & Communication in Aging and Neurodegeneration Research Group, led by Dr. Angela Roberts, at Northwestern University. The CHAT format included codings to indicate where and which errors and disfluencies occur [21]. We illustrate with an example sentence in Table 2. These transcripts also included the timings for the start and end of each sentence. We use these timings to segment the recordings and transcripts to the sentence-level. In all our analysis, we process sentence-level data and manually exclude off-script sentences, such as the participants asking questions.

Finally, clinicians assessed each participant's cognitive health using the Montreal Cognitive Assessment (MoCA) [22]. The MoCA score quantifies cognitive functioning ranging from 0 (low functioning) to 30 (healthy functioning). In this dataset, participants have MoCA scores of 26.5 ± 2.8.

3. Methodology

3.1. Counting errors and disfluencies

We evaluate three approaches for counting errors and disfluencies, where each approach involves a higher level of automation. In each approach, we extract three features: the **number of inserted words**, the **number of deleted words**, the **number of substituted words**. For the third paragraph, where words were intentionally transposed and transposition errors are common, we extract a fourth feature: the **number of transposed words**. Finally, for each paragraph, we extract the **number of pauses**. We extract these features at the sentence-level and aggregate them to the paragraph-level. We then analyze how errors and disfluencies from each paragraph perform in predicting CI for people with PD, and we explore how this changes across the three approaches for counting errors and disfluencies.

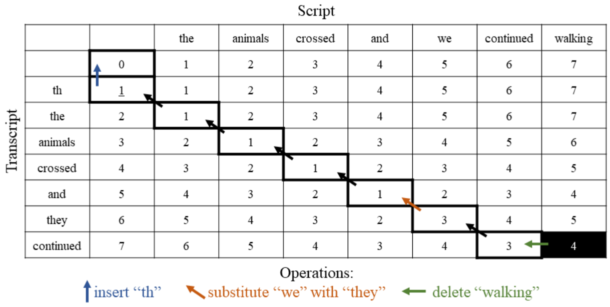
Manual CHAT features. We write a Python script to process CHAT format lines and extract error and disfluency features. We refer to these as the manual CHAT features.

Manual text features. We generate text versions of the manual transcripts by removing the CHAT codings. This version of the transcripts has only the words spoken by the participant. We provide an example in Table 2. We automatically extract error and disfluency features from these transcripts and we refer to these features as the manual text features. This analysis suggests that we may not need to rely on CHAT codings.

We use the Damerau-Levenshtein algorithm to compare each manual text transcript with its scripted sentence and count any language errors and disfluencies[1] [23, 24]. The Damerau-Levenshtein algorithm finds the minimum total cost of operations needed to transform one sentence into another, where valid operations include inserting, deleting, substituting, or transposing words. We use this algorithm to build a distance matrix between a transcribed and a scripted sentence, and we backtrack through this matrix to count the frequency of each operation. Figure 1 illustrates this process. As we build the distance matrix, we assign the following costs to each operation: insertion=1, deletion=1, substitution=2, and transposition=2. Note that substitutions and transpositions cost more because each could be accomplished with one insertion and one deletion. After building the distance matrix, there are sometimes several paths for backtracking and tallying operations. As we backtrack, we prioritize checking for transpositions, substitutions, insertions, and then deletions.

Script	the animals crossed and we continued walking
CHAT	↫th-the animals crossed and they [:: we] (.) continued &0walking
Text	th the animals crossed and they continued

Table 2: *Example script, CHAT transcript, and text transcript. In the CHAT transcript, specific codings indicate that the participant inserted "th," substituted "we" with "they," made one short pause, and deleted "walking." The text transcript has only the participant's speech.*



		the	animals	crossed	and	we	continued	walking
	0	1	2	3	4	5	6	7
th	1	1	2	3	4	5	6	7
the	2	1	2	3	4	5	6	7
animals	3	2	1	2	3	4	5	6
crossed	4	3	2	1	2	3	4	5
and	5	4	3	2	1	2	3	4
they	6	5	4	3	2	3	4	5
continued	7	6	5	4	3	4	3	4

Figure 1: *Example of distance matrix with backtracking. We first use dynamic programming to fill in the distance matrix. Then, starting from the bottom-right cell, we follow the lowest cost path to backtrack to the top-left cell. Each movement (up, to the left, or diagonally up and to the left) that is associated with a change in costs also indicates an edit operation.*

Finally, we use a Praat method introduced by de Jong and Wempe to count pauses [25, 26]. This method detects silences of at least 0.2 seconds where the volume drops 25 dB below the 99th percentile dB of the full audio file.

We compare the manual CHAT features to manual text features to evaluate the potential of automating error and disfluency detection given an audio recording and a transcript. In the "Manual Text" column of Table 4, we report the quadratic weighted kappa (QWK) [27], which quantifies agreement between the manual text and manual CHAT features. In many cases, specifically for insertions, deletions, and substitutions from the first and second paragraphs, the features from these two approaches are in near-perfect agreement (QWK=0.93-1.00). For these features from paragraph three, where the errors and disfluencies tend to be more complicated, there is slightly less agreement (QWK=0.82-0.99). There is also less agreement for pauses from these two approaches (QWK=0.19-0.62), highlighting the difficulty of detecting pauses.

ASR text features. Finally, we relax our dependence on manual transcription. We use the Mozilla DeepSpeech ASR system [28, 29] to investigate the feasibility of automating the transcription from which we can detect errors and disfluencies. This ASR system was built using LibriSpeech [30]. We fine-tune the acoustic model using our dataset. We build a tri-gram lan-

[1] https://github.com/amritkromana/reading-error-disfluency-features

Feature Type	Approach	Paragraph 1		Paragraph 2		Paragraph 3		All Paragraphs	
		CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE
Errors & Disfluencies	Manual CHAT	0.48	2.36	0.11	2.89	0.26	2.74	0.61	2.38
	Manual Text	0.35	2.58	0.12	2.84	0.59	2.21	0.64	2.30
	ASR Text	0.23	2.79	0.26	2.76	0.25	2.76	0.47	2.65
Acoustic	–	0.22	3.40	0.14	3.93	0.33	2.96	0.05	6.94

Table 3: *Results for predicting MoCA score with features from three paragraphs and errors and disfluencies counted with three different approaches. MoCA=Montreal Cognitive Assessment, CCC=concordance correlation coefficient, RMSE=root-mean-square error.*

Paragraph	Feature	Manual text	ASR text
1	Insertions	1.00	0.89
	Deletions	0.93	0.45
	Substitutions	0.97	0.42
	Pauses	0.19	0.19
2	Insertions	1.00	0.94
	Deletions	0.98	0.65
	Substitutions	0.98	0.55
	Pauses	0.45	0.45
3	Insertions	0.99	0.91
	Deletions	0.82	0.88
	Substitutions	0.87	0.60
	Transpositions	0.87	0.92
	Pauses	0.62	0.62

Table 4: *Quadratic weighted kappa, measuring agreement, between manual CHAT features and text features. Note: pause features are not dependent on text type.*

guage model [31] for each paragraph using the manual text transcripts. We perform both of these steps in a leave-one-subject-out (LOSO) manner, so the system does not have exposure to any of a participant's speech before generating their transcripts. The resulting ASR transcripts have a word error rate of 3.9%.

We extract features from the ASR transcripts and audio using the Damerau-Levenshtein and Praat methods described previously. We refer to these features as the ASR text features. Note that the pause features derive from the audio so these features are the same across the manual and ASR text feature sets.

We compare the manual CHAT features to ASR text features to evaluate the potential of fully automating the transcription and detection of errors and disfluencies. The "ASR Text" column in Table 4 details how the ASR text features agree with the manual CHAT features (QWK=0.42-0.94). However, these agreements are generally not as strong as those between the manual text features and manual CHAT features (QWK=0.82-1.00). This suggests the transcription errors made by the ASR system are propagating into the feature extraction.

3.2. Baseline acoustic features.

For comparison, we use Praat to extract a set of acoustic features that were analyzed in previous work [15]: **fundamental frequency** (standard deviation, relative standard deviation, variational range, relative variational range), **energy** (same four statistics), **total speech time**, and **speech index of rhythmicity**. Rektorova et al. demonstrated the relevance of these features to future cognitive decline in people with PD. However, it is unknown if these features could be predictors of CI at any one point in time. It is also unknown how these features may relate to CI when extracted from this unique reading task.

4. Results and discussion

We train a linear regression model to predict MoCA scores with different feature sets [32]. In all experiments, we train this model in a LOSO manner, ensuring that data from the test subject were not used to train the model, and we perform z-score normalization of the input features based on the training set. In our first experiment, we train this model with manual CHAT error and disfluency features. Next, we assess how our predictive ability changes as we move toward automated approaches with manual and ASR text features. We compare these results to a model trained with acoustic features. For each of these feature sets, we also compare using features from any single paragraph or using features from all three paragraphs. We report the condordance correlation coefficient (CCC) [33] and root-mean-square error (RMSE) between the predicted and ground truth MoCA scores. CCC quantifies the agreement between the predicted and ground truth scores while RMSE measures the magnitude of the difference. These results are listed in Table 3.

We find that using the manual CHAT error and disfluency features, the model's performance increases as we incorporate features from all paragraphs (CCC increases from 0.48 to 0.61, but RMSE also slightly increases from 2.36 to 2.38). Using the manual text features, the model's performance increases further (CCC=0.64 and RMSE=2.30). This improvement in performance results from small differences in how a manual annotator and the Damerau-Levenshtein method categorize errors and disfluencies. Using the ASR text features, the model still performs well (CCC=0.47 and RMSE=2.65), although we observe a drop in performance due to the ASR system not transcribing all errors and disfluencies. We find that a model trained with these error and disfluency features outperforms a model trained with acoustic features (CCC=0.33 and RMSE=2.96).

We analyze the learned beta coefficients from the models trained with error and disfluency features from each approach (Figure 2). We consider a feature important if its beta coefficient is significant (p-value < 0.05) for the majority of LOSO runs. The important manual CHAT error and disfluency features highlight the complex relationship between errors, disfluencies, and CI. In the first paragraph, people with PD who experience more CI (i.e. have lower MoCA scores) pause more ($\beta = -1.23 \pm 0.13$) and insert more words to repeat or revise their speech ($\beta = -0.90 \pm 0.08$), while people with PD who experience less CI (i.e. have higher MoCA scores) omit more words ($\beta = 1.29 \pm 0.08$). In the second paragraph, people with PD who experience more CI pause more ($\beta = -0.74 \pm 0.06$) and substitute more words ($\beta = -1.18 \pm 0.10$), while people with PD who experience less CI insert more words to correct their speech ($\beta = 1.20 \pm 0.11$). In the third paragraph, people with PD who experience more CI insert ($\beta = -1.13 \pm 0.09$) and substitute more words ($\beta = -0.94 \pm 0.09$). These errors and disfluencies may provide additional insight into how cognitive processes are impacted in people with PD.

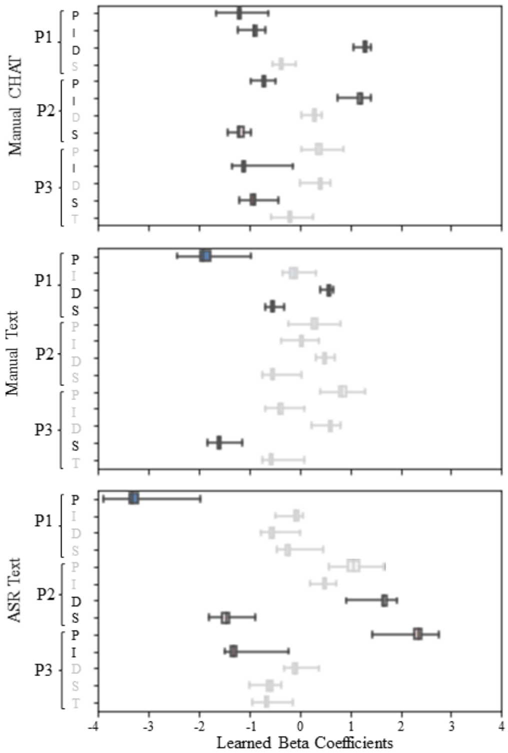
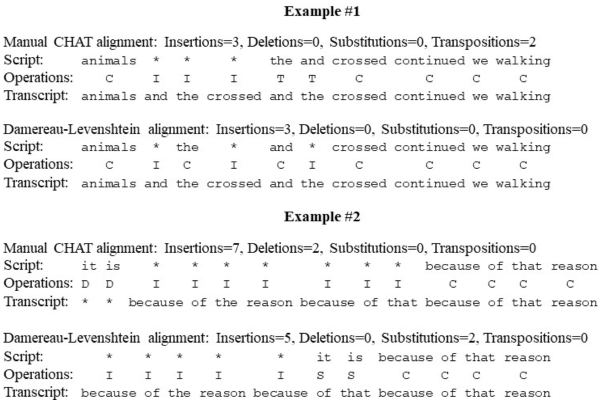


Figure 2: *Learned beta coefficients from MoCA prediction models using error features from three different approaches: manual CHAT, manual text, and ASR text. P1=first paragraph, P2=second paragraph, P3=third paragraph. P=pauses, I=insertions, D=deletions, S=substitutions, T=transpositions. The bold features are statistically significant (p-value < 0.05) in the majority of leave-one-subject-out model runs.*

As we move toward the more automated approaches, we find that there are changes in which features are important. The changes result from discrepancies in how the approaches detect and label errors and disfluencies. When errors and disfluencies are compounded, the Damerau-Levenshtein method detects and labels them to minimize the total operational cost, whereas a manual CHAT annotator may not. We illustrate some of the resulting alignment differences in Figure 3. Example #1 shows how the Damerau-Levenshtein method will not detect transpositions after a repeated phrase, because it can find these words in the correct order. Example #2 shows how the Damerau-Levenshtein method will always count a deletion followed by an insertion (or vice versa) as a substitution, even when a manual CHAT annotator would detect these errors separately. These disagreements, along with transcription errors, also drive the changes in feature importance for the ASR text features.

Despite these small discrepancies, our results demonstrate that these automated methods have considerable potential for transcribing and detecting errors and disfluencies. One feasible use case for these methods involves implementing them in a human-in-the-loop system. Rather than manually transcribing the audio, we can have a transcriber review and edit ASR transcripts. Given the audio and these transcripts, we can then use the Damerau-Levenshtein and Praat methods to propose a script-transcript alignment (such as those in Figure 3) with pauses. Instead of needing to manually detect all errors and disfluencies, a transcriber may need to just review these alignments for accuracy. Given an approved alignment, we can then automatically extract error and disfluency features and assess CI. This human-in-the-loop system may allow us to most quickly and accurately evaluate the extent of CI for treatment planning.



Example #1

```
Manual CHAT alignment: Insertions=3, Deletions=0, Substitutions=0, Transpositions=2
Script:       animals  *   *     *    the and crossed continued we walking
Operations:      C     I   I     I     T   T     C        C     C     C
Transcript:   animals and the crossed and the crossed continued we walking

Damereau-Levenshtein alignment: Insertions=3, Deletions=0, Substitutions=0, Transpositions=0
Script:       animals  *  the    *    and  *  crossed continued we walking
Operations:      C     I   C     I     C   I     C        C     C     C
Transcript:   animals and the crossed and the crossed continued we walking
```

Example #2

```
Manual CHAT alignment: Insertions=7, Deletions=2, Substitutions=0, Transpositions=0
Script:       it is    *    *   *    *       *    *   *   because of that reason
Operations:   D  D     I    I   I    I       I    I   I      C    C   C     C
Transcript:   *  *  because of the reason because of that because of that reason

Damereau-Levenshtein alignment: Insertions=5, Deletions=0, Substitutions=2, Transpositions=0
Script:          *    *   *    *       *    it  is  because of that reason
Operations:      I    I   I    I       I    S   S      C    C   C     C
Transcript:   because of the reason because of that because of that reason
```

Figure 3: *Examples in which the manual CHAT and Damerau-Levenshtein alignments differ. C=correct, I=insertion, D=deletion, S=substitution, T=transposition.*

5. Conclusions and future work

In this paper, we demonstrate the potential of using reading errors and disfluencies for predicting CI in people with PD. We first evaluate the use of error and disfluency features extracted from manual CHAT transcripts. We find that these features could be used in a regression framework to predict MoCA scores, where predicted and ground truth scores are correlated with CCC=0.61. We analyze the learned beta coefficients to understand the complex error and disfluency patterns resulting from CI. We then move to more automated approaches, finding that although these approaches have small differences in how they detect errors and disfluencies, compared to the manually extracted case, the features still predict CI (CCC=0.64 and 0.47). In future work, we will explore methods to bridge the gap between the manual and automated approaches for detecting and labeling errors and disfluencies. We will also investigate using more detailed labels, such as classifying insertions as revisions, repetitions, and filled pauses. Finally, we will analyze how errors and disfluencies change as a function of CI for people without PD. We hypothesize that these features will be sensitive to cognitive decline for people without PD, but the specific error and disfluency patterns may differ. The framework introduced in this work has the potential to enable this future analysis of how cognition modulates speech.

6. Acknowledgements

This work was funded in part by Precision Health at University of Michigan. Funding for the data was provided by a National Institute of Deafness and Communication Disorders (NIDCD) grant to Dr. Angela Roberts (NIH/NIDCD 1R21DC017255-01). Data collection and analysis were approved by and conducted in accordance with current human subjects ethics guidelines at Northwestern University (PI: Dr. Angela Roberts). We thank the following people who were involved in the manual aspects of the annotations: Stephanie Gutierrez, Brenda Xu, Abigael Parrish, and Richard Richter. Data were made available for this project through a data use agreement executed between Dr. Angela Roberts (Northwestern University) and Dr. Emily Mower Provost (University of Michigan).

7. References

[1] A. E. Lang and A. M. Lozano, "Parkinson's disease," *New England Journal of Medicine*, vol. 339, no. 16, pp. 1130–1143, 1998.

[2] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, and A. E. Lang, "Parkinson disease," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–21, 2017.

[3] C. C. Janvin, J. P. Larsen, D. Aarsland, and K. Hugdahl, "Subtypes of mild cognitive impairment in parkinson's disease: progression to dementia," *Movement disorders: official journal of the Movement Disorder Society*, vol. 21, no. 9, pp. 1343–1349, 2006.

[4] J. N. Caviness, E. Driver-Dunckley, D. J. Connor, M. N. Sabbagh, J. G. Hentz, B. Noble, V. G. H. Evidente, H. A. Shill, and C. H. Adler, "Defining mild cognitive impairment in parkinson's disease," *Movement disorders: official journal of the Movement Disorder Society*, vol. 22, no. 9, pp. 1272–1277, 2007.

[5] D. Aarsland, K. Bronnick, C. Williams-Gray, D. Weintraub, K. Marder, J. Kulisevsky, D. Burn, P. Barone, J. Pagonabarraga, L. Allcock *et al.*, "Mild cognitive impairment in parkinson disease: a multicenter pooled analysis," *Neurology*, vol. 75, no. 12, pp. 1062–1069, 2010.

[6] A. Gotham, R. Brown, and C. Marsden, "'frontal'cognitive function in patients with parkinson's disease 'on'and 'off'levodopa," *Brain*, vol. 111, no. 2, pp. 299–321, 1988.

[7] S. Skodda and U. Schlegel, "Speech rate and rhythm in parkinson's disease," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 7, pp. 985–992, 2008.

[8] A. Tsanas, M. Little, P. McSharry, and L. Ramig, "Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests," *Nature Precedings*, pp. 1–1, 2009.

[9] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson's disease symptom severity," *Journal of the royal society interface*, vol. 8, no. 59, pp. 842–855, 2011.

[10] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in parkinson's disease," *Journal of voice*, vol. 25, no. 4, pp. 467–472, 2011.

[11] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[12] J. Kim, M. Nasir, R. Gupta, M. V. Segbroeck, D. Bone, M. P. Black, Z. I. Skordilis, Z. Yang, P. G. Georgiou, and S. S. Narayanan, "Automatic estimation of parkinson's disease severity from diverse speech tasks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of parkinson's disease: a deep learning approach," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1618–1630, 2018.

[14] T. Thies, D. Mücke, A. Lowit, E. Kalbe, J. Steffen, and M. T. Barbe, "Prominence marking in parkinsonian speech and its correlation with motor performance and cognitive abilities," *Neuropsychologia*, vol. 137, p. 107306, 2020.

[15] I. Rektorova, J. Mekyska, E. Janousova, M. Kostalova, I. Eliasova, M. Mrackova, D. Berankova, T. Necasova, Z. Smekal, and R. Marecek, "Speech prosody impairment predicts cognitive decline in parkinson's disease," *Parkinsonism & related disorders*, vol. 29, pp. 90–95, 2016.

[16] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert *et al.*, "Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[17] T. Wang, C. Lian, J. Pan, Q. Yan, F. Zhu, M. L. Ng, L. Wang, and N. Yan, "Towards the speech features of mild cognitive impairment: Universal evidence from structured and unstructured connected speech of chinese." in *INTERSPEECH*, 2019, pp. 3880–3884.

[18] A. Garnham, R. C. Shillcock, G. D. Brown, A. I. Mill, and A. Cutler, "Slips of the tongue in the london-lund corpus of spontaneous conversation," in *Slips of the tongue and language production*. De Gruyter Mouton, 2011, pp. 251–264.

[19] T. H. Gollan and M. Goldrick, "Aging deficits in naturalistic speech production and monitoring revealed through reading aloud." *Psychology and aging*, vol. 34, no. 1, p. 25, 2019.

[20] S. Daniel and A. Lees, "Parkinson's disease society brain bank, london: overview and research." *Journal of neural transmission. Supplementum*, vol. 39, pp. 165–172, 1993.

[21] B. MacWhinney, *The CHILDES Project: Tools for analyzing talk. transcription format and programs*. Psychology Press, 2000, vol. 1.

[22] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[23] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.

[24] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[25] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[26] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[27] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[28] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[29] Mozilla, "Deepspeech," https://github.com/mozilla/DeepSpeech, 2017.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[33] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.