Cognitive and Acoustic Speech and Language Patterns Occurring in Different Neurodegenerative Disorders while Performing Neuropsychological Tests

M. Iglesias[1], A. Favaro[1], C. Motley[2], E. S. Oh[3,5], R.D. Stevens[4], A. Butala[2,5], L. Moro-Velázquez[1] and N. Dehak[1]

1. Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA
2. Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
3. Division of Geriatric Medicine and Gerontology, Dept. of Medicine, JHU SOM, Baltimore, MD, USA
4. Department of Anesthesiology and Critical Care, JHU SOM, Baltimore, MD, USA
5. Department of Psychiatry and Behavioral Sciences, Johns Hopkins Medicine, Baltimore MD, USA
{miglesi2, afavaro1, laureano, ndehak3}@jhu.edu

***Abstract*— In the last decade, improvements in automated speech processing, powered by signal processing and machine learning, has led to new approaches for medical assessment. Additionally, previous research in clinical speech has identified interpretable measures that are sensitive to changes in the cognitive, linguistic, affective, and motoric domains. In order to include speech-based automatic approaches in clinical applications, factors such as robustness, specificity, and interpretability of speech features are crucial. We focused on the analysis of a multi-modal array of interpretable features obtained from the spoken responses of participants with Neurodegenerative Diseases (ND) and control participants (CN) to neuropsychological tests. ND participants have Alzheimer's disease (AD), Parkinson's disease (PD), or Parkinson's disease mimics (PDM). We first collected spoken responses to three tests, a modified version of the Stroop test (MST), a verb naming task (VNT), and a noun naming task (NNT). Then, we arranged two complementary sets of cognitive and acoustic features and analyzed their statistical significance between the groups studied. Our results suggested that AD participants had significantly greater reaction times and significantly lower response accuracy with respect to the other groups across tests. In addition, PDM participants, compared to CN and PD participants, took a significantly longer time to complete the MST and NNT, while all the groups of participants with NDs showed significantly lower confidence during the MST. Since the analyzed features provided good differentiation results, they can be used in diagnostic tools for the assessment of NDs.**

***Keywords*—Alzheimer's disease (AD), Parkinson's disease (PD), biomarker, speech and language technologies, artificial intelligence**

I. Introduction

Neurodegenerative disorders (ND) are a group of neurological disorders with varying clinical features and pathological changes that affect specific subset of neurons, and are usually chronic and progressive [1]. Alzheimer's disease (AD) is the most common type of dementia, followed by Vascular Dementia and Dementia with Lewy Bodies (DLB) [2]. Parkinson's disease (PD) is also common, with millions of cases worldwide [3]. NDs vary considerably, in terms speed of onset, spectrum of resultant symptoms, or degree of impoverishment of quality of life. Some of them, such as AD and PD develop later in life, while others, such as cerebrovascular disease, migraines, and multiple sclerosis may develop across the age spectrum. Neurodegenerative disorders, such as AD and PD, are chronic and progressively worsen. This results in accumulating disability which may manifest with derangement in movement, behavior or cognition. At present, few diagnostic tools can reliably and easily distinguish NDs in their early stages without resorting to invasive procedures (e.g. lumbar punctures or repeated laboratory tests), or costly imaging studies. Moreover, the individual variability in symptom progression introduces additional challenges in accurate diagnosis. As changes in speech (e.g., diminished rate of speech or increase in quietness) are early indicators of NDS like amyotrophic lateral sclerosis and PD, speech-based biomarkers show potential as early detectors of NDs [4]. Recent advances in automated speech processing and machine learning techniques allow for the automatic extraction of speech-based biomarkers. Such features can be analyzed due the acoustic, articulatory, and linguistic information present in one's speech. For instance, Harel et al. [5] detected that for participants with PD, before a clinical diagnosis, fundamental frequency (F0) variability during free speech is diminished. Beltrami et al. [6] presented a set of features regarding lexical, acoustic and syntactic aspects that could distinguish between participants with multiple-domain mild cognitive impairment (MCI), elderly dementia, and healthy controls (CN). Other studies comparing individuals with dementia and CN [7] showed that individuals with dementia tended to have shorter response time and that differential frame-to-frame Jitter was the most significant distinguishing acoustic feature. The aforementioned studies were limited as they only analyzed linguistic and acoustic features derived from a singular language and task (i.e., usually a picture description task [8]) and had a narrow focus on one ND at a time. This work presents the analysis of a set of cognitive and acoustic features extracted from the spoken responses to three distinct neuropsychological tests: a modified Stroop test (MST), a noun naming task (NNT) and verb naming task (VNT). We employed different features to characterize participants' spoken responses and present a statistical analysis to examine the extent in which these features differ between participants with different NDs and CN. Our goal is to create automatic and objective methodologies utilizing different neuropsychological tests. These methodologies are intended to assist clinicians in assessing the presence and monitor the progression of different NDs. A summary of our automated pipeline is provided Fig 1.

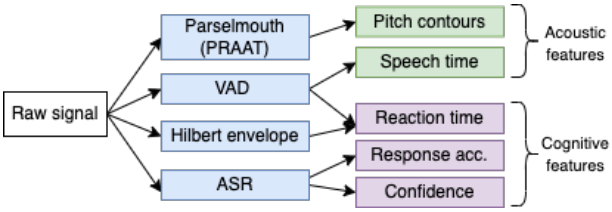

Figure 1. A block diagram of the main modules of our automated pipeline.

II. Materials

A. Data Collection

The data set, NeuroLogical Signals (NLS), is an ongoing data set collected by the authors of this study. It contains spoken responses to several tasks from participants with different NDs and CN participants. Clinical patients represent different NDs including, but not limited to, PD, AD, and Parkinson's-like diseases. Participants were also required to speak and read English fluently. All ND patients were seen at the Johns Hopkins Health System and all participants signed an informed consent document. This study was approved by the Johns Hopkins Medical Institutional Review Board. As this study was performed during the COVID-19 pandemic, all the participants wore the same surgical mask during recordings.

B. Participant Grouping

In the current study, we included individuals diagnosed with *clinically established* PD, AD, PD mimics (PDM) and CN participants. The PDM group was composed of people with a variety of presentations that are PD-like in nature. These included: Dementia with Lewy Bodies, Multiple System Atrophy, Tourette Syndrome, Dystonia, Spinocerebellar Ataxias and Wilsons disease. Prior to their final diagnosis, all PDM participants were diagnosed with possible PD. All diagnoses in the PDM group meet the highest clinical diagnostic criteria. Table I contains the baseline characteristics of the different ND groups.

C. Tasks

We analyzed spoken responses to three different neuropsychological tests. A trained research assistant gave instructions on how to perform each task prior to the start of each recording session. The first test was a modified version of the Stroop test. The *Stroop effect* measures how well someone overcomes the cognitive interference created by processing two stimuli at the same time, and can be tested via the Stroop test [9]. In our MST, only one word is displayed on the screen and participants were instructed to name the color of the word, not the word itself. The word displayed was either *red*, *blue*, or *green*, and the color was likewise red, blue or green. Presenting one word at a time allowed measurement of the time between stimulus and response. Participants were given an example prior to the test. The other two tasks belong to the Frontotemporal Lobar Degeneration (FTLD) Neuropsychological Test Module, specifically the Northwestern Naming Battery. Tasks from the FTLD-MOD neuropsychological battery are often used clinically, and for research purposes [10]. The first task was an NNT in which participants were given 4 seconds (s) to name an object displayed on a screen. The objects used were *glove*, *pepper*, *cat*, *apple*, *snake*, *suit*, *belt*, *scissors*, *socks*, and *elephant*. Participants were asked to name the object using only one word. The second task was a VNT that involves displaying a cartoon containing an action and asking the participant to name it in one word. Each image was displayed for 4 s. Participants were told their response could end either with or without the present particle *ing*. The actions used in this experiment were *zipping*, *barking*, *sweeping*, *spilling*, *throwing/tossing*, *praying*, *swimming*, *pouring*, *reading/studying*, and *crying*.
Each task was specifically chosen for their ability to measure ideally single word responses to a single image or word. Moreover, the tasks similarity to each other, especially the ONT and VNT, allows for comparison across tasks in addition to within the task.

III. Methods

A. Automatic Transcription

All recordings were automatically transcribed using a pre-trained conformer CTC model[1] for the Librispeech data set [11] built on top of icefall.[2] Since speakers with NDs tend to produce higher word error rate in automatic speech recognition [12], we manually refined any errant transcription.

B. Audio supervision

All the audios were supervised to ensure that they had appropriate quality. Criteria for appropriate quality includes understandable speech, minimal background noise, and a task related response. This level of quality was necessary for most feature analysis. Recordings with no response were kept. Some recordings that did not match the mentioned criteria, such as background noise and non-task related audio, were kept for the response accuracy analysis, as it is not impacted by these factors. The recordings without an automatically generated transcription were discarded, allowing for a greater degree of automation. The exception were three AD participants, which were transcribed manually to increase the sample size. Recordings were resampled from 24 kHz to 16 kHz as some of the employed signal processing libraries described in Section III-C required this sampling rate.

C. Feature Analysis

To analyze transcriptions and recordings from each task, we first arranged two complementary sets of cognitive and acoustic features. The features and the method adopted to calculate them are reported below. Some features were selected for their prior use in analyzing speech in a clinical setting. For example, F0 variability has been shown to change in patients with NDs, given that F0-related features can characterize prosody which can be affected by neurological impairment [5]. We later used

[1]https://huggingface.co/csukuangfj/icefall-asr-librispeech-conformer-ctc-jit-bpe-500-2021-11-09

[2]https://github.com/k2-fsa/icefall

TABLE I
DEMOGRAPHIC AND DISEASE SEVERITY STATISTICS OF THE STUDY POPULATION. WE REPORT SAMPLE SIZE, SEX, AGE DISTRIBUTION AND SCORES ON THE MONTREAL COGNITIVE ASSESSMENT (MOCA) FOR EACH EXPERIMENTAL GROUP. IN ADDITION, WE REPORT CLINICAL DEMENTIA RATING SCALE SUM OF BOXES (CDR-SB)FOR THE AD GROUP (NLS DATA SET) AND UNIFIED PARKINSON'S DISEASE RATING SCALE PART III (MDS-UPDRS III) FOR THE PD AND PDM GROUPS.

Category	Sample (n)			Age		MoCA		CDR-SB		MDS-UPDRS III	
	tot	female	male	avg	range	avg	range	avg	range	avg	range
CN	44	26	18	66.89	26-94	25.70	16-30	–	–	–	–
AD	11	2	9	70.00	58-84	19.45	6-30	4.3	1.5-14	–	–
PD	21	8	13	67.33	49-79	25.75	22-30	–	–	24.27	8-44
PDM	12	8	4	54.33	27-74	24.73	19-28	–	–	39.33	14-74

the cognitive and acoustic measurements to conduct a Kruskal-Wallis statistical test, with α = 0.05, in order to quantify the extent in which these features differed between the four participant's groups. The non-parametric Kruskal-Wallis test measures against a null hypothesis that the median ranks of the groups are equal. To account for False Discovery Rate, a Benjamini–Hochberg correction was applied [13].

1) Reaction Time

Reaction time (RT) was defined as the time elapsed between the presentation of a given stimulus (e.g., color word) and the participant's response. RT measures the time it takes to verbally react to the stimulus, not to respond correctly. RT was found using both a Hilbert envelope and two Voice Activity Detectors (VAD), the Cobra VAD (cVAD) [3] and the Silero VAD (sVAD)[4]. We employed two VADs to add more precision to our RT estimation. We first obtained the Hilbert envelope of the audio, which determined the start of any sound, voice and background alike. A Hilbert envelope allows for higher resolution, measuring the start of any audio to the millisecond. The two VADs have less resolution, as they utilize a frame shift larger than 30 ms but are able to distinguish between voices and background noises. After the generation of Hilbert RT (hRT) candidates the cVAD was used to generate possible RTs. The hRT was generated by determining when the value of the envelope exceeded the mean value of the envelope's most silent section by 50 standard deviations. Any hRT not within 160 milliseconds of a cVAD generated RT (cRT) was discarded. This was done to ensure that the hRT was associated to a voice, not any sound. Any hRT candidate that was not discarded was then compared to the sVAD RT (sRT). The hRT closest to the sRT was chosen as the final RT. If there was no sRT, the earliest hRT was selected. If there were no possible hRTs, the cRT closest to the sRT was selected. If there were no hRTs nor an sRT, the earliest cRT was chosen. If there was no possible hRT or cRT, the RT was registered as missing.

[3]https://github.com/Picovoice/cobra

[4]https://github.com/snakers4/silero-vad/

2) Response Accuracy

To determine the accuracy of the responses, we used the automatic transcriptions of the spoken responses collected during the three tasks under assessment. For the MST, a response was considered correct if the first task-related word was the color of the word. All other responses were marked as incorrect. For the NNT and VNT responses containing the right word at any time during the recording were deemed correct. This was done as many participants went beyond describing the image in one word, for example saying *boy throwing a ball* instead of *throwing*.

3) Confidence

To quantify the confidence of a participant's response to a given task, we measured the number of words contained in the response. The underlying idea behind the adoption of this feature is that the lower the number of circumlocutions and periphrasis that participants use in their answers, the higher the certainty they should have in recalling the target word during the task. We expect that participants with NDs would utter more words before recalling the correct one. We computed this feature using only the spoken responses collected during the MST since in the other two tasks participants were not explicitly instructed to adopt a single words to complete the tasks, as anticipated in Section II-C. To extract this feature from the speech transcripts we used the pretrained pipeline for English available on Spacy.[5]

4) Pitch Contours and Speech Time

To examine participants' prosodic abilities, we employed different techniques to perform an automatic analysis of the speech signal. Acoustic features such as the length of silent segments, fundamental frequency (F0) variability, and many others can be used to assess irregularities in the rhythm and timing of speech that often occur in motor and cognitive decay [14], [15]. Moreover, patients reporting right hemisphere (RH) stroke and people with selective NDs—e.g., PD, frontotemporal dementia, schizophrenia—may have trouble modulating their tone-of-voice to express sentence intonation and emotion [16].

[5]https://spacy.io/models/en#en_core_web_sm

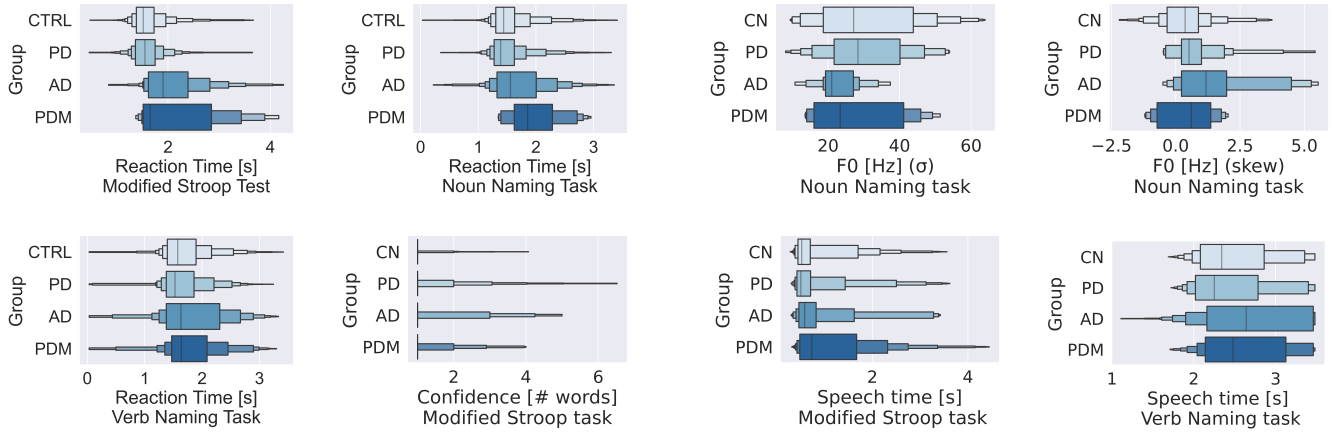

Figure 2. Categorical plots reporting some of the significant features ($p < 0.05$) from the cognitive domain.

We used Parselmouth, a Python library for the Praat software, to quantify prosody features, namely F0 standard deviation, skewness and kurtosis. To compute features based on F0 contours, we concatenated all the recordings that belong to a single speaker for each of the tasks under investigation. Moreover, we used DigiPsychProsody[6] to compute total speech time to assess whether participants with NDs took longer to complete the different tasks. This library uses the WebRTC Voice Activity Detector to create normalized features [7]. In total, 4 acoustic features were used. [8]

IV. Results and Discussion

We collected spoken responses to three different neurolopsychological tasks as indicated in Section III. We applied our extraction pipeline using the techniques introduced in Section III and Fig. 1. The results are summarized in Table II and commented in the sections below. For each significant comparison, we report results of the H-statistic, the corresponding p-value, the eta squared effect size (η^2) based on the H-statistic [17] and the area under the ROC curve (AUROC). AUROC can be used as a criterion to measure the feature's discriminative ability [18]. Features not reported in Table II were not significant in any pair. Though the groups are not perfectly age and gender-balanced, we do not expect that this affected our results in a significant way. The PDM, for example, is younger than the other three groups, but, compared to CNs, they under performed in most features.

A. Cognitive Features

The categorical plots reported in Fig. 2 represent the significant features ($p < 0.05$) from the cognitive domain. During the execution of the MST, all three groups with NDs showed a significantly lower level of confidence ($p < 0.05$) with respect to the CN group. This can be motivated by the need of inhibiting the cognitive interference originated during this task. Moreover,

[6]https://github.com/NeuroLexDiagnostics/DigiPsych_Prosody
[7]https://github.com/wiseman/py-webrtcvad
[8]https://github.com/wiseman/py-webrtcvad

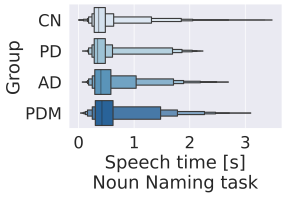

Figure 3. Categorical plots reporting some of the significant features ($p < 0.05$) from the acoustic domain.

results suggest that the AD group have significantly slower reaction time ($p < 0.05$) and less accurate responses ($p < 0.05$) in almost all the tasks under analysis. AD participants' accuracy was 67.7%, 86.9%, and 73.8% for the MST, NNT, and VNT respectively. The next lowest were 91.7%, 90.8%, and 88.1%, all from the PDM group. The PDM group was also characterized by a slower RT, even exceeding the AD group in the NNT. Studies on individuals with AD have shown that, compared to noun production, there exists a selective impairment in generating verbs [19], [20]. However, there are other studies that have shown that AD participants are impaired in both object and action naming, with a significantly larger deficiency in object naming [21]. Our experimental results seem to support findings from both of these two sides.

B. Acoustic Features

The categorical plots reported in Fig. 3 represent most of the significant features ($p < 0.05$) from the acoustic domain. During the execution of both the MST and the NNT, participants with PDM needed a significantly greater amount of time ($p < 0.05$) to complete the tasks. This phenomenon can be explained by the fact that half of participants with the PDM group were diagnosed with MCI, which is usually characterized by an early stage of memory loss and other cognitive ability (e.g., language, visual perception) [22]. Thus, the need to elaborate their responses to cognitive tasks may be the source of a longer speech time. For the AD group, results obtained for the cognitive features on the VNT seem to find confirmation in the acoustic domain. In fact, during this task participants with AD employed a significantly greater amount of time ($p < 0.05$) to

TABLE II

PAIRWISE KRUSKAL–WALLIS H TEST RESULTS FOR STATISTICALLY SIGNIFICANT FEATURES ($p < 0.05$) USING THE MST, VNT, AND NNT. FOR EACH PAIR-WISE COMPARISON, WE REPORT H-STATISTIC, THE CORRESPONDING p-VALUE, THE ETA SQUARED (η^2) EFFECT SIZE BASED ON THE H-STATISTIC COMPUTED AS $(H - k + 1)/(n - k)$, WHERE H IS THE VALUE OBTAINED IN THE KRUSKAL-WALLIS TEST; k IS THE NUMBER OF GROUPS; n IS THE TOTAL NUMBER OF OBSERVATIONS. WE ALSO REPORT THE AREA UNDER THE ROC CURVE (AUROC) FOR EACH OF THE SIGNIFICANT COMPARISONS.

Feature	**Sample** (n)		H	p-**value**	η^2	**AUROC**
	1	**2**				
Modified Stroop test						
Cognitive features						
Reaction time [s]	CN ($n = 41$)	AD ($n = 10$)	41.59	< 0.001	0.83	0.77
	PD ($n = 21$)	AD ($n = 10$)	35.21	< 0.001	1.18	0.78
Response accuracy	CN ($n = 41$)	AD ($n = 10$)	37.04	< 0.001	0.74	0.64
	PD ($n = 21$)	AD ($n = 10$)	21.66	< 0.001	0.71	0.63
	PD ($n = 21$)	PDM ($n = 12$)	11.18	0.002	0.33	0.62
Confindence [# words]	CN ($n = 41$)	AD ($n = 10$)	7.82	0.010	0.14	0.54
	CN ($n = 41$)	PDM ($n = 12$)	10.39	0.003	0.18	0.55
	CN ($n = 41$)	PD ($n = 21$)	5.77	0.03	0.08	0.53
Acoustic features						
Speech time [s]	CN ($n = 44$)	PDM ($n = 10$)	10.43	0.006	0.18	0.62
	PD ($n = 21$)	PDM ($n = 10$)	10.00	0.023	0.31	0.63
Verb Naming task						
Cognitive features						
Reaction time [s]	PDM ($n = 12$)	PD ($n = 19$)	6.73	0.02	0.20	0.59
Response accuracy	CN ($n = 41$)	AD ($n = 10$)	33.98	< 0.001	0.67	0.55
	PD ($n = 19$)	AD ($n = 10$)	12.81	< 0.001	0.44	0.55
	AD ($n = 10$)	PDM ($n = 12$)	7.50	0.01	0.33	0.52
Acoustic features						
Speech time [s]	CN ($n = 41$)	AD ($n = 10$)	7.87	0.020	0.14	0.59
	AD ($n = 10$)	PD ($n = 19$)	11.17	0.003	0.8	0.62
	PD ($n = 19$)	PDM ($n = 12$)	6.92	0.025	0.20	0.59
Noun Naming task						
Cognitive features						
Reaction time [s]	CN ($n = 41$)	AD ($n = 10$)	7.60	0.006	0.13	0.59
	CN ($n = 41$)	PDM ($n = 12$)	17.49	< 0.001	0.32	0.78
	PD ($n = 19$)	AD ($n = 10$)	10.17	0.001	0.34	0.61
	PDM ($n = 12$)	AD ($n = 10$)	5.95	0.03	0.25	0.32
	PDM ($n = 12$)	PD ($n = 19$)	19.23	< 0.001	0.63	0.19
Response accuracy	CN ($n = 41$)	AD ($n = 10$)	10.77	0.001	0.20	0.60
	PD ($n = 19$)	AD ($n = 10$)	10.27	0.001	0.34	0.58
	PD ($n = 19$)	PDM ($n = 12$)	5.03	0.049	0.14	0.60
Acoustic features						
F0 [Hz] (skew)	CN ($n = 41$)	AD ($n = 10$)	4.87	0.04	0.08	0.73
F0 [Hz] (kurt)	CN ($n = 41$)	AD ($n = 10$)	4.76	0.04	0.08	0.72
Speech time [s]	CN ($n = 41$)	PDM ($n = 12$)	8.72	0.017	0.15	0.59
	PD ($n = 19$)	PDM ($n = 12$)	6.90	0.023	0.20	0.59

deliver their responses. Moreover, once we concatenated all the spoken responses elicited during the 3 tasks under assessment, participants with AD turned out to have a lower F0 variability in the NNT.

V. CONCLUSION AND FUTURE WORK

In this work, we collected spoken responses to three neuropsychological tests and we adopted signal processing and machine learning techniques to arrange a multi-modal array of features to model the presence of different NDs along distinct dimensions. Our set of cognitive features represents valuable metrics to quantify the response time and confidence of the participants during the execution of the tasks well as the accuracy of their responses. On the other hand, the set of acoustic features encodes information about the time required by participants to complete a given task and fundamental frequency variability. All of these features were motivated by the clinical literature showing changes in voice and speech in participants with NDs. Overall results suggested that participants with AD had

significantly greater reaction times and significantly lower response accuracy with respect to the other experimental groups across tests. In particular, they exhibited greater difficulties during the execution of the VNT and NNT as anticipated in related studies. In addition, the AD group showed lower ability to modulate pitch. We expect to find a similar result for the PD group once we analyze tasks that require spontaneous speech elicited during a picture description, for instance. On the other hand, participants with PDM took significantly longer time to complete the MST and the NNT, while all the groups of participants with NDs showed a significantly lower confidence during the execution of the MST. In the future, we will include more participants to balance the experimental groups in terms of age, sex, and number of participants. Moreover, we will study the other tasks contained in our data set (i.e., reading passage, picture description) and will design new tasks to elicit different spoken responses. New tasks will test motor and memory abilities and capture the cognitive decline. On the whole, the ultimate goal of this study is developing a fully automated diagnostic pipeline that can help clinicians to perform precise assessment of the disorders and to monitor their progression in time, without the need of costly and time-consuming human analysis.

VI. Acknowledgements

This work was funded in part by the Richman Family Precision Medicine Center of Excellence – Venture Discovery Fund.

References

[1] World Health Organization et al., "Global action plan on the public health response to dementia 2017–2025," 2017.

[2] Serge Gauthier, Pedro Rosa-Neto, José Morais, and Claire Webster, "Journey through the diagnosis of dementia," *World Alzheimer Report 2021:*, , no. 1, 2021.

[3] Valery L Feigin, Emma Nichols, Tahiya Alam, Marlena S Bannick, Ettore Beghi, Natacha Blake, William J Culpepper, E Ray Dorsey, Alexis Elbaz, Richard G Ellenbogen, et al., "Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 5, pp. 459–480, 2019.

[4] Michael Neumann, Oliver Roesler, Jackson Liscombe, Hardik Kothare, David Suendermann-Oeft, David Pautler, Indu Navar, Aria Anvar, Jochen Kumm, Raquel Norel, et al., "Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale," *arXiv preprint arXiv:2104.07310*, 2021.

[5] Brian T Harel, Michael S Cannizzaro, Henrí Cohen, Nicole Reilly, and Peter J Snyder, "Acoustic characteristics of parkinsonian speech: a potential biomarker of early disease progression and treatment," *Journal of Neurolinguistics*, vol. 17, no. 6, pp. 439–453, 2004.

[6] Daniela Beltrami, Gloria Gagliardi, Rema Rossini Favretti, Enrico Ghidoni, Fabio Tamburini, and Laura Calzà, "Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?," *Frontiers in aging neuroscience*, vol. 10, pp. 369, 2018.

[7] Honghuang Lin, Cody Karjadi, Ting FA Ang, Joshi Prajakta, Chelsea McManus, Tuka W Alhanai, James Glass, and Rhoda Au, "Identification of digital voice biomarkers for cognitive health," *Exploration of medicine*, vol. 1, pp. 406, 2020.

[8] Harold Goodglass, Edith Kaplan, and Sandra Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*, Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[9] Federica Scarpina and Sofia Tagini, "The stroop color and word test," *Frontiers in psychology*, vol. 8, pp. 557, 2017.

[10] Cynthia K Thompson, Sladjana Lukic, Monique C King, M Marsel Mesulam, and Sandra Weintraub, "Verb and noun deficits in stroke-induced and primary progressive aphasia: The northwestern naming battery," *Aphasiology*, vol. 26, no. 5, pp. 632–655, 2012.

[11] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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.

[12] Laureano Moro-Velazquez, JaeJin Cho, Shinji Watanabe, Mark A Hasegawa-Johnson, Odette Scharenborg, Heejin Kim, and Najim Dehak, "Study of the performance of automatic speech recognition systems in speakers with parkinson's disease," *Proc. Interspeech 2019*, pp. 3875–3879, 2019.

[13] Yosef Hochberg and Yoav Benjamini, "More powerful procedures for multiple significance testing," *Statistics in medicine*, vol. 9, no. 7, pp. 811–818, 1990.

[14] Yasir Tahir, Debsubhra Chakraborty, Justin Dauwels, Nadia Thalmann, Daniel Thalmann, and Jimmy Lee, "Non-verbal speech analysis of interviews with schizophrenic patients," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5810–5814.

[15] Steven H Ferris and Martin Farlow, "Language impairment in alzheimer's disease and benefits of acetylcholinesterase inhibitors," *Clinical interventions in aging*, vol. 8, pp. 1007, 2013.

[16] Sona Patel, Kenichi Oishi, Amy Wright, Harry Sutherland-Foggio, Sadhvi Saxena, Shannon M Sheppard, and Argye E Hillis, "Right hemisphere regions critical for expression of emotion through prosody," *Frontiers in neurology*, vol. 9, pp. 224, 2018.

[17] Maciej Tomczak and Ewa Tomczak, "The need to report effect size estimates revisited. an overview of some recommended measures of effect size," *Trends in sport sciences*, vol. 1, no. 21, pp. 19–25, 2014.

[18] Suzanne Ekelund, "Roc curves—what are they and how are they used?," *Point of care*, vol. 11, no. 1, pp. 16–21, 2012.

[19] Mikyong Kim and Cynthia K Thompson, "Verb deficits in alzheimer's disease and agrammatism: Implications for lexical organization," *Brain and language*, vol. 88, no. 1, pp. 1–20, 2004.

[20] Keith M Robinson, Murray Grossman, Tammy White-Devine, and Mark D'Esposito, "Category-specific difficulty naming with verbs in alzheimer's disease," *Neurology*, vol. 47, no. 1, pp. 178–182, 1996.

[21] David JG Williamson, John C Adair, Anastasia M Raymer, and Kenneth M Heilman, "Object and action naming in alzheimer's disease," *Cortex*, vol. 34, no. 4, pp. 601–610, 1998.

[22] Serge Gauthier, Barry Reisberg, Michael Zaudig, Ronald C Petersen, Karen Ritchie, Karl Broich, Sylvie Belleville, Henry Brodaty, David Bennett, Howard Chertkow, et al., "Mild cognitive impairment," *The lancet*, vol. 367, no. 9518, pp. 1262–1270, 2006.