High frequency post-disfluency word choices and task-dependent speech behavior characterize connected speech in individuals with mild cognitive impairment

Michael J. Kleiman, PhD[1*] and James E. Galvin, MD, MPH[1]

[1] Comprehensive Center for Brain Health, Department of Neurology, University of Miami Miller School of Medicine, Boca Raton, FL 33433

Running Title: Post-disfluency speech behavior in MCI (39 characters)
Title: 18 words, 154 characters (with spaces)
Abstract: 242 words (250 max)
Text: 4446 words
Tables: 6
Figures: 3
References: 60
Total Pages (including references, figure, figure legends, and tables): 33

***Corresponding Author:**
Michael J. Kleiman, PhD
Comprehensive Center for Brain Health
University of Miami Miller School of Medicine
7700 W Camino Real, Suite 200
Boca Raton, FL 33433
E-mail: mjkleiman@miami.edu

ABSTRACT

Background: Alzheimer's disease (AD) is characterized by progressive cognitive decline, including impairments in speech production and fluency. Mild cognitive impairment (MCI), a prodrome of AD, has also been linked with changes in speech behavior but to a more subtle degree.

Objective: This study aimed to investigate whether speech behavior immediately following disfluencies (post-disfluency speech behavior) differs between individuals with MCI and healthy controls (HCs), and how these differences are influenced by the cognitive demands of various speech tasks.

Methods: Transcribed speech samples were analyzed from both groups across different tasks, including immediate and delayed narrative recall, picture descriptions, and free responses. Key metrics including lexical and syntactic complexity, both overall and post-disfluency, were examined.

Results: Significant differences were observed between the MCI and HC groups in terms of filled disfluencies, particularly "uh", with a higher incidence and longer latencies following these disfluencies in the MCI group. Word frequency following filled disfluencies was higher among MCI participants, with this effect varying based on the cognitive load of the tasks. Predictive analyses utilizing random forest classifiers demonstrated high specificity in using speech behavior metrics to differentiate between MCI and HCs.

Conclusions: Speech behavior following disfluencies differs between MCI participants and healthy controls, with these differences being influenced by the cognitive demands of the speech tasks. Post-disfluency speech metrics can be easily integrated into existing speech analysis

paradigms. This study contributes to the growing evidence on the significance of nuanced speech analysis in understanding and predicting cognitive decline in MCI.

Keywords: Mild cognitive impairment, speech behavior, neuropsychological assessment, machine learning

INTRODUCTION

Alzheimer's disease (AD) is characterized by progressive degradation of cognitive abilities that interfere with everyday functioning. While episodic memory is often impacted early, other domains also may show early deficits. The regions associated with language and speech production are often among the earliest affected [1,2]. Recent studies have provided insight into how AD affects speech production, including the increased use of more common words [3,4], a potentially compensatory mechanism of declining executive functioning, and less content-filled language [5] which may reflect degradation of semantic memory. Individuals with AD have also been found to exhibit increased pauses during speech [5–7], potentially indicative of difficulties in planning and increased lexical search.

Mild cognitive impairment (MCI), a prodromal stage of AD [8], has been associated with the degradation of more subtle measures of speech behavior in its earliest stages. Measures of speech fluency, including rate of speech and syllable count, have been shown to significantly degrade in MCI [9,10]. These are generally measures of connected speech performance; a form of speech in which each individual word is "connected" lexically to the next, as in a passage or narrative. One of the most common methods of obtaining a sample of connected speech is directing the participant to describe a scene or image, such as the Cookie Theft image [11], which enables participants to produce a descriptive sample of speech that is also restricted to a particular topic to promote scoring ability. A more open-ended approach is to simply ask the participant to freely answer a given question, such as "what did you do yesterday?", which prompts the production of more natural and conversational speech behavior. When scoring episodic memory is more important than capturing naturalistic speech, a paragraph recall task may also be used, as in the Wechler Logical Memory or Craft Story tests [12,13]. In these tasks, a story is read or shown to a participant

who is then asked to recall as many details about the story as possible both immediately following the story, as well as after a delay.

In connected speech tasks, some studies find that MCI is associated with more common, less complex language [14–16], which may be related to a delay in accessing lexical information [17] resulting in a tendency to use high frequency (more common) words, however other studies do not observe this increase in word frequency within MCI participants [18]. This trend towards common language is most visible in more severe forms of dementia, including semantic dementia [19], primary progressive aphasia [20], and AD [3,4], suggesting that while the limited findings of increased word frequency due to MCI may be reflective of markers of future decline, the heterogeneity of the disorder may make it more difficult for these markers to be detected.

Disfluencies, or periods between clauses where a speaker has paused, have also been shown to be affected in MCI. Disfluencies are characterized as either filled or unfilled. While unfilled disfluencies consist of extended pauses between words without any vocalization, filled disfluencies include utterances such as “uh”, “um”, and “er”, and are often used when a speaker is attempting to think of the next word or phrase, marking a region in which the speaker is engaging in word retrieval or language planning [21]. The usage of different “forms” of filled disfluencies have also been shown to vary in their usage; “uh” is typically used to indicate a short pause, while “um” is more often followed by longer pauses [22]. The frequency of these disfluency forms differ based on the task administered [23] which matches findings showing that disfluency usage in MCI changes based on the type of assessment; MCI is associated with increases in only filled disfluencies in free-response tasks [24,25] and either only unfilled disfluencies [25] or both filled and unfilled disfluencies [26] in narrative recall tasks.

Given that disfluencies are often produced during word searching behavior [21], and impaired word searching behavior has been found to lead to increased word frequency in MCI [14,16,17], it seems likely that speech behavior immediately following disfluencies ("post-disfluency" behavior) is affected in some way. Thus, we hypothesize that words immediately following disfluencies are more likely to be higher frequency and syntactically simple in MCI participants compared to post-disfluency word choices in cognitively-intact healthy controls (HCs). Further, previous studies have shown that speech behavior is variably affected by task [16,25,27], as well as by the cognitive demands involved in performing that task [10]. As a result, we expect that differences in disfluencies and post-disfluency word choices will differ based on the degree of task demands for each of the three tasks that we will examine: narrative recall (immediate and delayed), picture descriptions, and free response. We thus secondarily hypothesize that not all disfluencies are lexically-driven; increased cognitive load during more difficult tasks (as in the delayed narrative recall) may increase disfluencies but without affecting post-disfluency word frequency, and vice versa.

METHODS

Participants

53 participants (39 HCs and 14 participants with MCI) were selected from the Healthy Brain Initiative, a longitudinal study of brain health and cognition at the University of Miami Comprehensive Center [28] based on their consensus clinical diagnosis (no cognitive impairment or mild cognitive impairment) and their age (greater than 60 years old). A full description of the study protocol is described elsewhere [28], but in-brief every participant undergoes identical comprehensive clinical, cognitive, functional, and behavioral assessments, including the Clinical Dementia Rating (CDR) [29], a complete physical and neurological examination, and

neuropsychological test battery from the Uniform Data Set [30]. Healthy controls are defined as individuals with a global CDR of 0, unimpaired activities of daily living, and normal age- and education-normed neuropsychological test performance. Individuals with MCI were defined as a global CDR 0.5, unimpaired activities of daily living, and neuropsychological test performance in at least 1 domain greater than 1.5 standard deviations below the norm. Imaging and plasma biomarker collection is ongoing so only clinical diagnoses were considered here. Exclusion criteria included presence of clinically diagnosed aphasia, schizophrenia, generalized anxiety disorder or major depressive disorder, and non-English speakers. All protocols were approved by the Institutional Review Board at the University of Miami Miller School of Medicine.

Equipment and Materials

Participants were seated in a quiet area with minimal visual and auditory distractions in front of a 24" computer monitor attached to a Windows 10 machine. Participants interacted with the machine using custom-developed software written in Python and C, as well as a backlit keypad for indicating intention (finished with response, need help, etc.) and a volume knob. Participants wore Sony WH1000XM4 noise-cancelling headphones to further reduce auditory distractions. The software displayed visual and auditory stimuli for all tasks as well as written and audio instructions. Participants were given the opportunity to indicate whether they did not understand instructions and needed further explanation, after which an experimenter would assist them.

Procedure

Speech Tasks

Three speech-focused tasks were administered to participants: narrative recall with delay, picture description, and a spontaneous free response. Participant audio was recorded using a hyper-cardioid microphone to minimize background noise.

The **narrative recall (NR)** task uses a custom-developed story ("Maddy the dog") structured to be similar to the Craft Story 21 and Weschler Logical Memory narratives, with 25 content units of varying complexity [31] and developed to be used alongside the Craft Story 21 [13] which is administered to all participants in a separate visit as part of their longitudinal evaluations. This story differs from traditional narrative assessment in that it is both verbally read and visually presented to participants, promoting stronger multimodal encoding than simple auditory presentation. Additionally, the delayed portion was collected between fifteen and twenty minutes after initial presentation of the story, in contrast to the traditional 30 minutes, to both streamline data collection and explore differences between the already collected Craft Story delay. During this delay, additional speech tasks were administered. Participants were given two minutes for both the immediate and delayed responses.

Picture descriptions (PD) for a black and white version of the Modern Cookie Theft image [32] were collected from each participant, chosen as it is a highly validated and commonly-used stimulus for eliciting descriptive spontaneous speech. The updated image, as opposed to the original line-drawing version [11], is a more modern depiction of a kitchen scene which avoids stereotypical gender roles, adds additional content units, and incorporates more robust shading and coloring as opposed to the flat line drawings of the original which may aid participants in visually exploring the scene [33]. This updated version has been shown to elicit more detailed descriptions even under time constraints [34]. We further transformed this version by desaturating the image while still maintaining high contrast, as preliminary study revealed that many participants overly

relied on color-based descriptions and less on describing action. Participants are given ninety seconds to describe each image, with a visual indicator at the 80-second mark. Additionally, if they choose to end their description before the time limit, they are asked "is there anything else?".

Spontaneous free responses (FR) to the prompt "Describe your typical morning routine" were also collected. Participants were given no additional prompting, with their responses recorded until they indicated completion or until 90 seconds had passed.

Transcripts

Audio recordings from each task were processed using *OpenAI Whisper*'s Large-v2 model [35], a speech-to-text model that generates transcripts at higher accuracy than many other free and paid options and is capable of being run on in-house hardware. Each automatically generated transcript was then manually corrected as needed by trained research staff. Generated transcripts included start and stop times in seconds for each word as well as confidence for each predicted word which helped highlight areas of potentially necessary correction.

Audio Preprocessing

All audio files were preprocessed prior to speech-to-text transcription. Leading and trailing silence was removed using *PyDub* [36], and the end of a file was marked using a chime to encourage *OpenAI Whisper* to end sentences appropriately when subjects ran out of time and the audio file ended abruptly. If a chime is not used, hallucinations where *Whisper* finished sentences were common, including erroneously completed sentences with predicted text or an ending message (e.g., "The end.") Background noise, including room tone and non-target extraneous speech (e.g., voices from hallway) were removed using *PySoX* [37] to further improve speech-to-text accuracy as well as to facilitate audiometric analysis.

Analysis

Analysis of Speech Transcripts

After speech responses were transcribed and validated, the transcripts were parsed using a Python script incorporating multiple NLP packages, including *SpaCy* for tokenization, lemmatization, part of speech analysis, and general NLP analyses [38], *wordfreq* for examining word frequency using its own proprietary formula [39], and *textdescriptives* for generating a wide range of statistics including entropy, perplexity, sentence count and length, syllable count, and reading ease statistics such as Flesch Kincaid grade level, Gunning-Fog, and the Coleman Liau Index [40,41]. Custom scripts were also written that generate SUBTLEX rankings [42] for determining word frequency, mean Yngve depth for examining passage complexity [43], and type token ratios and derivatives (MATTR or "moving average type token ratio" [44] and MTLD or "measure of textual diversity" [45]) for lexical diversity.

Disfluencies were defined as either filled or unfilled, with unfilled disfluencies defined as pauses lasting greater than 0.75 seconds between words, and filled disfluencies defined as the presence of "um", "uh", and "er". The four words immediately following the disfluency were marked as "post-disfluency" words, and statistics were generated for both the average of all four words and only the first word following the disfluency. These disfluency and post-disfluency statistics included total counts and durations of disfluencies, latencies, syllable counts, SUBTLEX rankings, *wordfreq* frequencies, and part of speech frequencies.

Statistical and Predictive Analyses

Transcripts were compared between groups for both individual and combined tasks using the *pingouin v0.5.3* [46] Python package. Analyses of covariance (ANCOVA) with age as a covariate

were performed for all measures where assumptions of normality and homoscedasticity were met, and Kruskal-Wallis Rank Sum test when assumptions were not met and age was not significantly different between groups as determined by linear regression. In cases where assumptions were not met and age was a significant factor, comparisons were not considered. Statistical logistic regressions were also used to determine the combined effects of multiple variables, especially when isolating the effects of post-disfluency metrics alone.

Predictive analyses, including random forest classifiers, gradient boosted machines, logistic regressions, and support vector classifiers, were performed to identify the predictive power of post-disfluency and general speech metrics to determine binary impairment status (impaired vs. not impaired). All analyses were cross-validated using a repeated stratified K-folds procedure (3-fold, 3-repeat), which resulted in nine combinations of train/test sets for better generalizability of model results. Outputs were evaluated using sensitivity, specificity, F1 score, and area under the ROC curve (AUC). The *LightGBM v4.2.0* [47] Python package's implementation was used for gradient boosted machines, and all other predictive analyses and methods were performed using *scikit-learn v1.3.2* [48].

Two sets of features were used in our models; one including all statistically significant features, and another including only statistically significant post-disfluency metrics. Only individual task metrics were used; when significant results for were found in the mixed ANOVA containing all tasks, only the task with the highest Bayes index in the post-hoc was used in an effort to minimize the negative effect of multicollinearity on model results. Additionally, feature selection was performed using the *BorutaSHAP v1.1* [49] selection algorithm using a random forest classifier (*scikit-learn v1.3.0* [48]) as its base model. One-fourth (25%) of the stratified training data was held for use in the feature selection process, to avoid leakage and subsequent overfitting in the

model training phase. *Optuna v3.5.0* [50] selected optimal hyperparameters for each model, leveraging its implementation of define-by-run dynamic parameter search spaces and efficient strategies for pruning, using the same 25% of stratified training data in the feature selection step.

RESULTS

Sample Characteristics

There was a significant difference in age between the two groups (p = .003), resulting in age being included as a covariate in all following comparisons (**Table 1**). No differences between groups were found for gender, years of education, race or ethnicity, vulnerability [51], or resilience [52].

Common Findings in All Tasks

When examining all tasks using a Mixed ANOVA, with task as the within-subjects variable and impairment status as the between-subjects variable, differences were found between the HC and MCI groups. There were significantly more filled disfluencies ("um", "uh", and "er") in the MCI group (4.21 ± 4.42) compared to the HC group (2.47 ± 2.71), p=.006, especially the "uh" disfluency (2.75±3.53 MCI vs 1.11±1.78 HC; $\chi^2(1)=13.31$, p<.001), with latencies between all disfluencies and the following word increasing in the MCI group (1.22±0.35 MCI vs 1.11±0.33 HC; $\chi^2(1)=9.31$, p=.002), **Figure 1**.

For post-disfluency metrics, the Kruskal-Wallis non-parametric ANOVA test was required due to many participants not producing disfluencies at all in some tasks, resulting in unequal variances between groups. Word frequency was significantly higher following a filled disfluency in MCI participants, supporting our hypotheses; *wordfreq* rankings following "uh" disfluencies (.0087±.013 MCI vs .0058±.012 HC; $\chi^2(1)=7.31$, p=.007) were significantly different. Latencies

following "um" disfluencies were also significantly higher in MCI participants (1.83 sec ± 0.73 MCI vs 1.33 sec ± 0.56 HC; $\chi^2(1)=17.36$, $p<.001$).

Task Comparisons

All differences between tasks described below exhibited a significant ANOVA or pairwise t-test within-subjects score, with an alpha set at .01.

Picture description

In the PD task, the mean Yngve depth was shallower in MCI participants (3.36±0.67 levels) than for HC participants (3.98±0.86 levels), p=.012. The Coleman-Liau Index, a measure of readability, also decreased in MCI participants (3.65±1.45 MCI vs 4.74±0.88 HC; p=.008) (**Table 2.c**).

Narrative recall

In both narrative recall tasks, filled disfluencies significantly increased in MCI participants (3.23±3.97 disfluencies MCI vs 1.58±2.28 HC; $\chi^2(1)=7.37$, p=.007), however this significance did not appear in pairwise comparisons (**Table 2.a, 2.b**). Latencies following "um" disfluencies increased in MCI participants in both NR tasks (1.82±0.72 sec MCI vs 1.25±0.49 sec HC; $\chi^2(1)=8.43$, p=.004), however post-hoc analyses revealed that only the delayed NR task showed significance between groups (**Table 2.b**).

Mean total word counts decreased in both immediate and delayed NR tasks (74.75±30.21 words in MCI vs 92.88±31.95 words in HC; p=.002), with lexical diversity measured by MTLD found to decrease in only the immediate NR task (38.79±16.53 MCI vs 50.42±15.64 HC; p=.013). Also in the immediate NR task, the Coleman Liau Index decreased from 6.74±1.67 in HC to 5.27±2.31 in MCI, p=.007. Mean word lengths (3.94±0.26 characters MCI vs 4.16±0.24 characters HC; p=.003)

in the immediate NR task also decreased in MCI participants, while median word length decreased in the delayed NR task (3.50±0.48 MCI vs 3.90±0.47 HC; p=.002). The number of syllables per word in the immediate NR task was also found to decrease for MCI participants (1.12±0.06 syllables MCI vs 1.16±0.05 syllables HC; *p*=.006). Fewer nouns were used by MCI participants in the immediate NR task (9.82±3.23 nouns MCI vs 12.73±3.77 nouns HC; *p*=.007). Further, proper nouns such as names were significantly less common for MCI participants in both the immediate NR (2.71±2.39 words MCI vs 5.24±2.54 words HC; *p*=.001) and delayed NR tasks (1.94±1.69 words MCI vs 4.89±2.92 words HC; *p*<.001).

Free response

The only metric significant in the free response task were word frequencies as calculated by *wordfreq,* which were significantly more common in MCI participants after "uh" disfluencies in this task (.0086±.0078 MCI vs .0025±.0027 HC; *p*=.005) (**Table 2.d**), but not in other tasks (**Figure 2**).

Predictive analysis

When starting with all significant features, feature selection narrowed the field to 16 of the best-performing features: total disfluency count (immediate NR), total filled disfluency count (immediate NR), total "uh" disfluency count (immediate and delayed NR), mean *wordfreq* frequency post-"uh" (FR), mean latency post-"um" (delayed NR), word count (immediate NR), MTLD (immediate NR), mean *wordfreq* frequency (immediate NR), median word length (delayed NR) Coleman-Liau Index (PD & immediate NR), mean Yngve depth (PD), proper noun ratio (immediate NR), and noun count (delayed NR). Using this set of features and after performing

hyperparameter optimization, the best performing model was the random forest classifier with a diagnostics odds ratio (DOR) of 13.52 (**Table 3**) and an AUC of 0.828 (**Figure 3)**.

When using only the two most significant post-disfluency metrics (word frequency following "uh" disfluencies in the FR task, and latencies following "um" disfluencies in the delayed NR task), the best performing model was again the random forest classifier with a DOR of 10.3 (**Table 3**) and an AUC of 0.791 (**Figure 3)**.

DISCUSSION

In this study, we examined whether post-disfluency speech is affected due to cognitive impairment, and if task demands elicit differences in speech behavior between HC and MCI participants. We found significant differences between groups in word frequency following filled disfluencies in the FR task, as well as increased latencies following "um" disfluencies in the NR tasks.

Task-dependent disfluency production

MCI participants tend to use more common (high frequency) language following filled "uh" disfluencies, but only in the FR task. Our initial hypothesis that lexical frequency would decrease in MCI individuals due to increased word searching behavior may have underestimated the intervening effect of task demands. Just as different forms of filled disfluencies differ in their usage [22] it is possible that each form may be produced for different reasons as task conditions vary. The question in the FR task asked about the participant's typical routine, which is a topic that is typically static, repetitive, and familiar, requiring little cognitive load to recall. Disfluencies produced in this context likely resemble typical conversational speech, and would match the documented usage of "uh" disfluencies as indicators of word-searching behavior [22,53,54]. As a result, cognitive impairment would play a larger role in disrupting this search, leading to greater

differences between the HC and MCI groups. This is contrasted with the three other tasks, where required cognitive demands are greater. In these tasks, “uh” disfluencies may still be produced as indicators of searching behavior, but instead of lexical search displayed in the spontaneous speech task, this searching behavior may be more directed towards searching for the correct answer (in NR tasks) or searching for a new object to describe (in the PD task). Thus, while post-disfluency lexical frequency may still be affected by cognitive impairment, the HC group may also use more common words as a result of the increased cognitive load, resulting in a less noticeable difference between groups.

In a similar vein, the delayed NR task was arguably the most difficult task administered in this study, requiring the most cognitive demands to recall information presented after being distracted in the interim with other demanding speech tasks. This may explain why post-“um” latencies were significantly different between groups only in this task. “Um” disfluencies have been shown to signal longer delays than other disfluencies [22,54], and the delays decrease with skill [21] and increase with cognitive load [55]. Delayed narrative recall performance is known to be significantly affected due to cognitive impairment, and so MCI participants likely needed to utilize more cognitive resources to perform the task, leading to increased delays compared to the HC group.

While previous studies have shown significant differences between MCI and HC in numbers of disfluencies, both filled and unfilled [25,26] and especially using the “uh” utterance [56], our results did not show significant differences between the two groups in total count of disfluencies when examining the post-hoc pairwise tests for task comparison that were not also significantly different due to age effects. While our sample size is relatively limited, the observation of significant age effects may suggest that previous identification of increased numbers of

287 disfluencies in these studies may be due to the MCI group being older than the HC group. Future
288 examination into total disfluency counts should take age into account as a possible covariate.

289 *Other lexical differences between tasks*

290 More significant differences between groups were found in the immediate NR task than any others,
291 and when multiple tasks exhibited the same differences in a single measure the *Boruta* feature
292 selection algorithm identified the immediate NR task as the most predictive. The MTLD measure
293 of lexical diversity as well as the Coleman-Liau index both identify a tendency of MCI participants
294 to use simpler language in this task, as does the finding of reduced mean word length and syllables
295 per word. The high usefulness of the immediate NR task has been previously shown in prior work
296 developing parsimonious assessment composites for detecting MCI [57], where it was determined
297 to be one of the most useful metrics for identifying MCI among the neuropsychological tasks
298 administered by ADNI. This may be due to this task requiring coordination of working memory
299 and attention, two areas known to be degraded in MCI but not entirely disrupted.

300 Proper noun usage decreased in both the immediate and delayed NR tasks, indicating that while
301 healthy participants are able to recall names and places, those with MCI are less likely to recall
302 this information. As both were impaired, it is difficult to determine whether the deficit was due to
303 encoding or retrieval, however previous study by Mueller et al. [58] showing that proper noun
304 recall in delayed but not immediate NR is associated with preclinical AD suggests that retrieval is
305 the main factor affected in proper noun recall.

306 Mean Yngve depth, a measure of syntactic complexity based on hierarchical relationships between
307 clauses in produced speech, was found to be significantly shallower in picture descriptions of the
308 Cookie Theft image produced by MCI participants. Picture description tasks elicit hierarchical

language as participants describe the relationships between objects in a picture or photograph. The Cookie Theft picture in particular contains many relationships between objects, with for example the boy engaging with both the cookie jar and the stool and the man both washing the dishes and not attending to them resulting in them overflowing. The reduction of mean Yngve depth in MCI may a result of a reduced focus on how objects in a scene relate to each other, instead resulting in simpler descriptions of these objects. Previous studies have shown that cognitive impairment affects mean Yngve depth in NR tasks [20,59], but PD tasks have not been previously shown to elicit decreased syntactic complexity. In our sample, mean Yngve depth did not appear to differ between groups for either NR task but did differ in the PD task. It is possible that the updated Cookie Theft image utilized in our study [32] encouraged either increased interrelated descriptions in healthy controls or a decrease in MCI participants. The PD task also prompted decreased word lengths, Coleman-Liau indices, and noun counts in MCI participants, indicating less precise and simpler language in this task, with this finding supported by previous studies [14,60].

Use of speech behavior to predict impairment status

After performing feature selection on the entire group of available features, a random forest utilizing the 16 selected features produced a model with an AUC of 0.828 and a DOR of 13.52. While specificity was excellent (94.59%), sensitivity for detecting impairment was only 43.59%. Nonetheless, if the task is to screen for impairment, the use of speech-based markers alone perform well at reducing numbers of non-impaired individuals. When only the top two post-disfluency features are used, only marginal decreases in sensitivity and specificity are observed. With high specificity and AUC, the results of this model suggest that post-disfluency metrics of latency and word frequency would be an excellent addition to any machine learning model that utilizes speech behavior and seeks to categorize healthy individuals in some capacity.

Limitations

The main limitations of this study were the limited sample size and the lack of biomarker data available. As MCI is a heterogeneous disorder, it is possible that separation of MCI subtypes (e.g., MCI due to AD, MCI due to Lewy body or vascular etiologies) will lead to more detailed characterization using neurobehavioral markers such as speech. Additional recruitment and biomarker examination of existing and new participants is ongoing.

We also observed age differences between groups. Instead of reducing our healthy control sample to minimize age effects through age-matching, we chose to instead include age as a covariate within our analyses due to the small size of the impaired group. Our future research will incorporate age-matching, as we aim to recruit a sufficient number of biomarker-confirmed impaired participants to support an evenly sized control group.

CONCLUSIONS

Reduced post-disfluency word frequency in the absence of task demands and increased post-disfluency latency in tasks with high cognitive load were observed in MCI participants, with both able to accurately predict impairment status with an AUC of 79.1%. Future research will examine likely causes of disfluency production, comparing lexical or word-finding disfluencies with those driven by task demands or cognitive load, as well as neural correlates of speech degradation.

ACKNOWLEDGEMENTS

Work on this study was supported by grants from the National Institute on Aging (R01 AG071514, R01 AG069765, and R01 NS101483), the Alzheimer's Association (AARF-22-923592), the

Evelyn F. McKnight Brain Research Foundation (FP00006751), and the Harry T. Mangurian Foundation.

DISCLOSURE STATEMENT

The authors have no conflict of interest to report.

REFERENCES

[1] Forbes-McKay KE, Venneri A. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. Neurol Sci 2005;26:243–54. https://doi.org/10.1007/s10072-005-0467-9.

[2] Ahmed S, Haigh A-MF, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain 2013;136:3727–37. https://doi.org/10.1093/brain/awt269.

[3] Kavé G, Goral M. Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. J Clin Exp Neuropsychol 2016;38:958–66. https://doi.org/10.1080/13803395.2016.1179266.

[4] Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease 2015;49:407–22. https://doi.org/10.3233/JAD-150520.

[5] Yamada Y, Shinkawa K, Nemoto M, Ota M, Nemoto K, Arai T. Speech and language characteristics differentiate Alzheimer's disease and dementia with Lewy bodies. Alz & Dem Diag Ass & Dis Mo 2022;14. https://doi.org/10.1002/dad2.12364.

[6] Yuan J, Cai X, Bian Y, Ye Z, Church K. Pauses for Detection of Alzheimer's Disease. Front Comput Sci 2021;0. https://doi.org/10.3389/fcomp.2020.624488.

[7] Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected Speech in Neurodegenerative Language Disorders: A Review. Frontiers in Psychology 2017;8:269. https://doi.org/10.3389/fpsyg.2017.00269.

[8] Petersen RC. Mild Cognitive Impairment. Continuum (Minneap Minn) 2016;22:404–18. https://doi.org/10.1212/CON.0000000000000313.

[9] Themistocleous C, Eckerström M, Kokkinakis D. Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. PLOS ONE 2020;15:e0236009. https://doi.org/10.1371/journal.pone.0236009.

[10] Mueller KD, Koscik RL, Hermann BP, Johnson SC, Turkstra LS. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. Frontiers in Aging Neuroscience 2018;9:1–14. https://doi.org/10.3389/fnagi.2017.00437.

[11] Goodglass H, Kaplan E, Weintraub S. BDAE: The boston diagnostic aphasia examination. Lippincott Williams & Wilkins Philadelphia, PA; 2001.

[12] Wechsler D. Wechsler memory scale-revised. The psychological corporation. New York 1987.

[13] Craft S, Newcomer J, Kanne S, Dagogo-Jack S, Cryer P, Sheline Y, et al. Memory improvement following induced hyperinsulinemia in alzheimer's disease. Neurobiology of Aging 1996;17:123–30. https://doi.org/10.1016/0197-4580(95)02002-0.

[14] Beltrami D, Gagliardi G, Rossini Favretti R, Ghidoni E, Tamburini F, Calzà L. Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline? Front Aging Neurosci 2018;0. https://doi.org/10.3389/fnagi.2018.00369.

[15] Roark B, Mitchell M, Hollingshead K. Syntactic complexity measures for detecting mild cognitive impairment. Proceedings of the Workshop on BioNLP 2007 Biological, Translational, and Clinical Language Processing - BioNLP '07, Prague, Czech Republic: Association for Computational Linguistics; 2007, p. 1. https://doi.org/10.3115/1572392.1572394.

[16] Clarke N, Barrick TR, Garrard P. A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning. Frontiers in Computer Science 2021;3.
[17] Taler V, Kousaie S, Sheppard C. Lexical access in mild cognitive impairment. The Mental Lexicon 635775264000000000;10:271–85. https://doi.org/10.1075/ml.10.2.05tal.
[18] Fraser KC, Lundholm Fors K, Eckerström M, Öhman F, Kokkinakis D. Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers. Frontiers in Aging Neuroscience 2019;11:1–18. https://doi.org/10.3389/fnagi.2019.00205.
[19] Meteyard L, Quain E, Patterson K. Ever decreasing circles: Speech production in semantic dementia. Cortex 2014;55:17–29. https://doi.org/10.1016/j.cortex.2013.02.013.
[20] Fraser KC, Meltzer JA, Graham NL, Leonard C, Hirst G, Black SE, et al. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex 2014;55:43–60. https://doi.org/10.1016/j.cortex.2012.12.006.
[21] Womack K, McCoy W, Ovesdotter Alm C, Calvelli C, Pelz JB, Shi P, et al. Disfluencies as Extra-Propositional Indicators of Cognitive Processing. In: Morante R, Sporleder C, editors. Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics, Jeju, Republic of Korea: Association for Computational Linguistics; 2012, p. 1–9.
[22] Clark HH, Fox Tree JE. Using uh and um in spontaneous speaking. Cognition 2002;84:73–111. https://doi.org/10.1016/S0010-0277(02)00017-3.
[23] Shriberg E. To 'errrr' is human: ecology and acoustics of speech disfluencies. Journal of the International Phonetic Association 2001;31:153–69. https://doi.org/10.1017/S0025100301001128.
[24] Toth L, Hoffmann I, Gosztolya G, Vincze V, Szatloczki G, Banreti Z, et al. A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. Current Alzheimer Research 2017;14:130–8. https://doi.org/10.2174/1567205014666171121114930.
[25] Vincze V, Szatlóczki G, Tóth L, Gosztolya G, Pákáski M, Hoffmann I, et al. Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild Alzheimer's disease. Clinical Linguistics & Phonetics 2020;00:1–16. https://doi.org/10.1080/02699206.2020.1827043.
[26] Egas-López JV, Balogh R, Imre N, Hoffmann I, Szabó MK, Tóth L, et al. Automatic screening of mild cognitive impairment and Alzheimer's disease by means of posterior-thresholding hesitation representation. Computer Speech & Language 2022:101377. https://doi.org/10.1016/j.csl.2022.101377.
[27] Wang T, Hong Y, Wang Q, Su R, Ng M, Xu J, et al. Identification of Mild Cognitive Impairment Among Chinese Based on Multiple Spoken Tasks. Journal of Alzheimer's Disease 2021;82:1–20. https://doi.org/10.3233/JAD-201387.
[28] Besser LM, Chrisphonte S, Kleiman MJ, O'Shea D, Rosenfeld A, Tolea M, et al. The Healthy Brain Initiative (HBI): A prospective cohort study protocol. PLOS ONE 2023;18:e0293634. https://doi.org/10.1371/journal.pone.0293634.
[29] Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology 1993;43:2412–2412. https://doi.org/10.1212/WNL.43.11.2412-a.
[30] Weintraub S, Besser L, Dodge HH, Teylan M, Ferris S, Goldstein FC, et al. Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set

(UDS). Alzheimer Disease and Associated Disorders 2018;32:10–7. https://doi.org/10.1097/WAD.0000000000000223.
[31] Bolognani SAP, Miranda MC, Martins M, Rzezak P, Bueno OFA, de Camargo CHP, et al. Development of alternative versions of the Logical Memory subtest of the WMS-R for use in Brazil. Dement Neuropsychol 2015;9:136–48. https://doi.org/10.1590/1980-57642015DN92000008.
[32] Berube S, Nonnemacher J, Demsky C, Glenn S, Saxena S, Wright A, et al. Stealing Cookies in the Twenty-First Century: Measures of Spoken Narrative in Healthy Versus Speakers With Aphasia. Am J Speech Lang Pathol 2019;28:321–9. https://doi.org/10.1044/2018_AJSLP-17-0131.
[33] Heuer S. The influence of image characteristics on image recognition: a comparison of photographs and line drawings. Aphasiology 2016;30:943–61. https://doi.org/10.1080/02687038.2015.1081138.
[34] Hux K, Frodsham K. Speech and Language Characteristics of Neurologically Healthy Adults When Describing the Modern Cookie Theft Picture: Mixing the New With the Old. Am J Speech Lang Pathol 2023;32:1110–30. https://doi.org/10.1044/2022_AJSLP-22-00291.
[35] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision 2022. https://doi.org/10.48550/arXiv.2212.04356.
[36] Robert J, Webbie M, Jean-philippe S, Ramdasan A, Lee C, Campbell J, et al. PyDub 2023.
[37] Bittner R, Humphrey EJ, Bello J. PySOX: Leveraging the Audio Signal Processing Power of SOX in Python. New York, NY, USA: Proceedings of the 17th International Society for Music Information Retrieval Conference Late Breaking and Demo Papers; 2016.
[38] Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: Industrial-strength Natural Language Processing in Python 2020. https://doi.org/10.5281/ZENODO.1212303.
[39] Speer R. rspeer/wordfreq: v3.0 2022. https://doi.org/10.5281/zenodo.7199437.
[40] Hansen L, Olsen LR, Enevoldsen K. TextDescriptives: A Python package for calculating alarge variety of metrics from text. JOSS 2023;8:5153. https://doi.org/10.21105/joss.05153.
[41] Coleman M, Liau TL. A computer readability formula designed for machine scoring. Journal of Applied Psychology 1975;60:283–4. https://doi.org/10.1037/h0076540.
[42] Brysbaert M, New B. Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav Res Methods 2009;41:977–90. https://doi.org/10.3758/BRM.41.4.977.
[43] Yngve VH. A Model and an Hypothesis for Language Structure. Proceedings of the American Philosophical Society 1960;104:444–66.
[44] Covington MA, McFall JD. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). Journal of Quantitative Linguistics 2010;17:94–100. https://doi.org/10.1080/09296171003643098.
[45] McCarthy PM, Jarvis S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods 2010;42:381–92. https://doi.org/10.3758/BRM.42.2.381.
[46] Vallat R. Pingouin: statistics in Python. JOSS 2018;3:1026. https://doi.org/10.21105/joss.01026.

[47] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc.; 2017.
[48] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–30.
[49] Keany E. BorutaShap: A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. 2020. https://doi.org/10.5281/zenodo.4247618.
[50] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA: Association for Computing Machinery; 2019, p. 2623–31. https://doi.org/10.1145/3292500.3330701.
[51] Kleiman MJ, Galvin JE. The Vulnerability Index: A weighted measure of dementia and cognitive impairment risk. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 2021;13:e12249. https://doi.org/10.1002/dad2.12249.
[52] Galvin JE, Kleiman MJ, Chrisphonte S, Cohen I, Disla S, Galvin CB, et al. The Resilience Index: A Quantifiable Measure of Brain Health and Risk of Cognitive Impairment and Dementia. Journal of Alzheimer's Disease 2021;84:1729–46. https://doi.org/10.3233/JAD-215077.
[53] Swerts M. Filled pauses as markers of discourse structure. Journal of Pragmatics 1998;30:485–96. https://doi.org/10.1016/S0378-2166(98)00014-9.
[54] Degand L, Gilquin G, Meurant L, Simon AC. Fluency and Disfluency Across Languages and Language Varieties. Presses universitaires de Louvain; 2019.
[55] Russo M, Bendazzoli C, Defrancq B. Making Way in Corpus-based Interpreting Studies. Springer; 2017.
[56] Yuan J, Bian Y, Cai X, Huang J, Ye Z, Church K. Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. Interspeech 2020, ISCA; 2020, p. 2162–6. https://doi.org/10.21437/Interspeech.2020-2516.
[57] Kleiman MJ, Barenholtz E, Galvin JE. Screening for Early-Stage Alzheimer's Disease Using Optimized Feature Sets and Machine Learning. Journal of Alzheimer's Disease 2021;81:355–66. https://doi.org/10.3233/JAD-201377.
[58] Mueller KD, Koscik RL, Du L, Bruno D, Jonaitis EM, Koscik AZ, et al. Proper names from story recall are associated with beta-amyloid in cognitively unimpaired adults at risk for Alzheimer's disease. Cortex 2020;131:137–50. https://doi.org/10.1016/j.cortex.2020.07.008.
[59] Roark B, Mitchell M, Hosom J-P, Hollingshead K, Kaye J. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. IEEE Transactions on Audio, Speech, and Language Processing 2011;19:2081–90. https://doi.org/10.1109/TASL.2011.2112351.
[60] Mueller KD, Hermann B, Mecollari J, Turkstra LS. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. Journal of Clinical and Experimental Neuropsychology 2018;40:917–39. https://doi.org/10.1080/13803395.2018.1446513.

Table 1. Subject characteristics

	Healthy Controls	**Mild Cognitive Impairment**	**p-val**
Age	68.28 (7.21)	76.64 (8.90)	**0.000**
% Female	0.72 (0.45)	0.64 (0.50)	0.478
Years of Education	16.43 (2.65)	17.00 (3.22)	0.920
% Non-White	0.11 (0.31)	0.21 (0.43)	0.159
% Hispanic	0.12 (0.33)	0.21 (0.43)	0.130
MoCA Total	27.05 (2.26)	23.29 (2.73)	**0.000**
FAQ Total	0.14 (0.45)	0.18 (0.60)	0.750
Vulnerability Index	6.55 (2.02)	8.64 (2.73)	0.078
Resilience Index	183.33 (33.06)	172.54 (32.30)	0.326
Number Symbol Coding	45.89 (9.16)	31.36 (6.49)	**0.000**

All comparisons performed using ANCOVA with age as a covariate, except age which used ANOVA.
Bold = significant

Table 2.a. Speech measures for the Immediate Narrative Recall task

Speech measure	HC	MCI	Stat (df = 60)	p-val	Test
Total Disfluency Count	3.390 (2.509)	6.067 (3.770)	8.469	0.004[X]	KW
Unfilled Disfluency Count	2.073 (1.694)	2.533 (1.407)	1.578	0.209	KW
"Uh" Disfluency Count	0.707 (1.470)	1.333 (1.633)	3.167	0.075	KW
"Um" Disfluency Count	0.610 (1.321)	2.200 (3.570)	2.509	0.113	KW
Post-"uh" Word Frequency	0.008 (0.016)	0.004 (0.008)	0.029	0.866	KW
Post-"um" Mean Latency	1.395 (0.471)	1.762 (0.721)	1.970	0.160	KW
Total Word Count	84.378 (26.173)	70.059 (24.517)	3.818	0.055	ANOVA
MTLD	50.422 (15.642)	38.791 (16.528)	6.616	**0.013**	ANOVA
Mean Word Frequency	0.006 (0.001)	0.007 (0.002)	0.557	0.458	ANOVA
Mean Word Length	4.158 (0.239)	3.944 (0.258)	9.449	**0.003**	ANOVA
Median Word Length	3.967 (0.165)	3.735 (0.437)	7.772	0.005[X]	KW
Mean Syllables per Word	1.156 (0.046)	1.117 (0.055)	7.980	**0.006**	ANOVA
Coleman-Liau Index	6.737 (1.670)	5.271 (2.306)	7.660	**0.007**	ANOVA
Mean Yngve Depth	4.785 (1.450)	4.739 (1.733)	0.011	0.917	ANOVA
Noun Count	12.733 (3.774)	9.824 (3.226)	7.902	**0.007**	ANOVA
Proper Noun Count	5.244 (2.542)	2.706 (2.392)	12.694	**0.001**	ANOVA

KW = Kruskal-Wallis; ANOVA = Analysis of Variance; HC = Healthy Control; MCI = Mild cognitive impairment; MTLD = measure of textual lexical diversity; X = Significant age effects
Bold = significant

Table 2.b. Speech measures for the Delayed Narrative Recall task

Speech measure	HC	MCI	F (df = 60)	p-val	Test
Total Disfluency Count	4.558 (3.712)	4.938 (3.623)	0.024	0.877	KW
Unfilled Disfluency Count	2.721 (3.850)	1.812 (1.223)	0.206	0.650	KW
“Uh” Disfluency Count	1.140 (1.656)	1.062 (1.611)	0.123	0.726	KW
“Um” Disfluency Count	0.698 (1.264)	2.062 (2.720)	4.682	0.030	KW
Post-“uh” Word Frequency	0.003 (0.003)	0.005 (0.004)	1.231	0.267	KW
Post-“um” Mean Latency	1.174 (0.500)	1.886 (0.766)	5.740	**0.015**	KW
Total Word Count	96.065 (35.668)	79.750 (35.422)	2.492	0.120	ANOVA
MTLD	50.375 (23.418)	45.202 (14.668)	0.683	0.412	ANOVA
Mean Word Frequency	0.007 (0.002)	0.007 (0.001)	0.146	0.704	ANOVA
Mean Word Length	4.109 (0.334)	3.889 (0.226)	5.954	0.018	ANOVA
Median Word Length	3.902 (0.467)	3.500 (0.483)	9.827	**0.002**	KW
Mean Syllables per Word	1.168 (0.054)	1.137 (0.060)	3.747	0.058	ANOVA
Coleman-Liau Index	5.905 (2.242)	5.128 (2.519)	1.339	0.252	ANOVA
Mean Yngve Depth	4.639 (1.404)	5.156 (1.741)	1.420	0.238	ANOVA
Noun Count	15.848 (5.692)	12.688 (6.488)	3.405	0.070	ANOVA
Proper Noun Count	4.891 (2.915)	1.938 (1.692)	13.026	**0.000**	ANOVA

KW = Kruskal-Wallis; ANOVA = Analysis of Variance; HC = Healthy Control; MCI = Mild cognitive impairment; MTLD = measure of textual lexical diversity; X = Significant age effects
Bold = significant

Table 2.c. Speech measures for the Picture Description task

Speech measure	HC	MCI	F (df = 60)	p-val	Test
Total Disfluency Count	5.600 (2.988)	7.294 (2.756)	1.188	0.280[X]	ANCOVA
Unfilled Disfluency Count	2.978 (2.572)	3.176 (1.976)	0.083	0.775	ANOVA
"Uh" Disfluency Count	1.178 (1.482)	1.471 (1.068)	0.553	0.460	ANOVA
"Um" Disfluency Count	1.400 (1.959)	2.588 (2.476)	3.916	0.052	ANOVA
Post-"uh" Word Frequency	0.010 (0.016)	0.014 (0.019)	1.720	0.190	KW
Post-"um" Mean Latency	1.498 (0.720)	1.912 (0.752)	2.658	0.103	KW
Total Word Count	134.356 (55.009)	114.176 (43.805)	1.840	0.180	ANOVA
MTLD	38.641 (11.760)	33.141 (12.431)	2.617	0.111	ANOVA
Mean Word Frequency	0.010 (0.002)	0.010 (0.002)	0.934	0.338	ANOVA
Mean Word Length	3.981 (0.143)	3.843 (0.179)	4.638	0.035[X]	ANCOVA
Median Word Length	3.356 (0.472)	3.000 (0.000)	8.613	**0.003**	KW
Mean Syllables per Word	1.217 (0.040)	1.191 (0.053)	0.337	0.564[X]	ANCOVA
Coleman-Liau Index	4.740 (0.884)	3.651 (1.448)	7.069	**0.008**	KW
Mean Yngve Depth	3.983 (0.863)	3.356 (0.668)	4.913	**0.012**	ANOVA
Noun Count	32.933 (13.048)	25.941 (9.959)	3.987	0.05	ANOVA
Proper Noun Count	0.002 (0.006)	0.001 (0.003)	0.268	0.606	ANOVA

KW = Kruskal-Wallis; ANOVA = Analysis of Variance; HC = Healthy Control; MCI = Mild cognitive impairment; MTLD = measure of textual lexical diversity; X = Significant age effects
Bold = significant

Table 2.d. Speech measures for the Free Response task

Speech measure	HC	MCI	F (df = 60)	p-val	Test
Total Disfluency Count	6.477 (3.944)	8.200 (6.201)	0.430	0.512[X]	KW
Unfilled Disfluency Count	2.864 (2.758)	2.067 (2.086)	1.089	0.297	KW
"Uh" Disfluency Count	1.932 (2.366)	1.933 (1.981)	0.105	0.746	KW
"Um" Disfluency Count	1.682 (2.197)	4.200 (4.945)	3.281	0.070[X]	KW
Post-"uh" Word Frequency	0.003 (0.003)	0.009 (0.008)	8.027	**0.005**	KW
Post-"um" Mean Latency	1.252 (0.391)	1.731 (0.716)	5.250	0.022	KW
Total Word Count	126.000 (71.155)	119.938 (84.409)	0.078	0.781	ANOVA
MTLD	42.294 (24.467)	41.056 (21.867)	0.032	0.859	ANOVA
Mean Word Frequency	0.007 (0.001)	0.007 (0.002)	1.177	0.282	ANOVA
Mean Word Length	3.597 (0.260)	3.572 (0.212)	0.121	0.729	ANOVA
Median Word Length	3.130 (0.341)	3.031 (0.125)	0.000	0.992[X]	ANCOVA
Mean Syllables per Word	1.152 (0.057)	1.145 (0.043)	0.207	0.650	ANOVA
Coleman-Liau Index	3.446 (1.797)	3.530 (1.239)	0.030	0.864	ANOVA
Mean Yngve Depth	4.867 (1.909)	4.731 (1.369)	0.068	0.795	ANOVA
Noun Count	24.130 (13.016)	19.375 (11.949)	1.649	0.204	ANOVA
Proper Noun Count	0.009 (0.016)	0.012 (0.032)	0.122	0.728	ANOVA

KW = Kruskal-Wallis; ANOVA = Analysis of Variance; HC = Healthy Control; MCI = Mild cognitive impairment; MTLD = measure of textual lexical diversity; X = Significant age effects
Bold = significant

Table 3. Measures of predictive ability for models utilizing speech behavior

	Sensitivity	Specificity	PPV	NPV	DOR	AUC
All Significant Features	43.59%	94.59%	73.91%	82.68%	13.52	0.828
Post-Disfluency Features	41.03%	93.70%	69.57%	81.89%	10.33	0.791

All Significant Features: Total disfluencies (INR), Total filled disfluencies (INR), Total "uh" disfluencies (INR & DNR), Mean *wordfreq* frequency post- "uh" (FR), mean latency post-"um" (DNR), word count (INR), MTLD (INR), mean *wordfreq* frequency (INR), median word length (DNR) Coleman-Liau Index (PD & INR), mean Yngve depth (PD), proper noun ratio (INR), noun count (DNR)
Post-disfluency Features: Mean *wordfreq* frequency post- "uh" (FR), mean latency post-"um" (DNR)

FR="Free-response task"; INR="Immediate Narrative Recall task"; DNR="Delayed Narrative Recall task"; PD="Picture Description task"; PPV="Positive Predictive Value"; NPV="Negative Predictive Value"; DOR="Diagnostic Odds Ratio"; AUC="Area under the ROC curve"

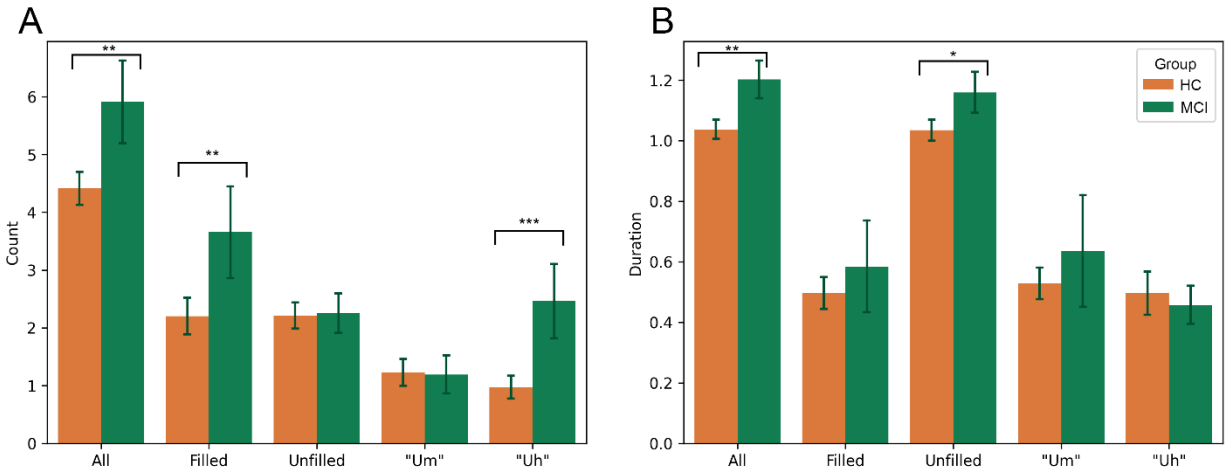

Figure 1. Count and duration of disfluency types separated by impairment status. (A) Total number of disfluencies separated by type of disfluency (all disfluencies, filled, unfilled, filled with "um", filled with "uh") and by impairment status (HC in orange vs MCI in green). All tasks were included in these comparisons. Significant differences between impairment status can be observed when examining All disfluencies, Filled disfluencies, and "Uh" disfluencies, all of which increased in the MCI group, with no differences found for unfilled and "um" disfluencies when all tasks were included. (B) Total duration in seconds separated by type of disfluency and impairment status. Unfilled disfluencies and All disfluencies were found to significantly increase in the MCI group. Significance is indicated using asterisks, with *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

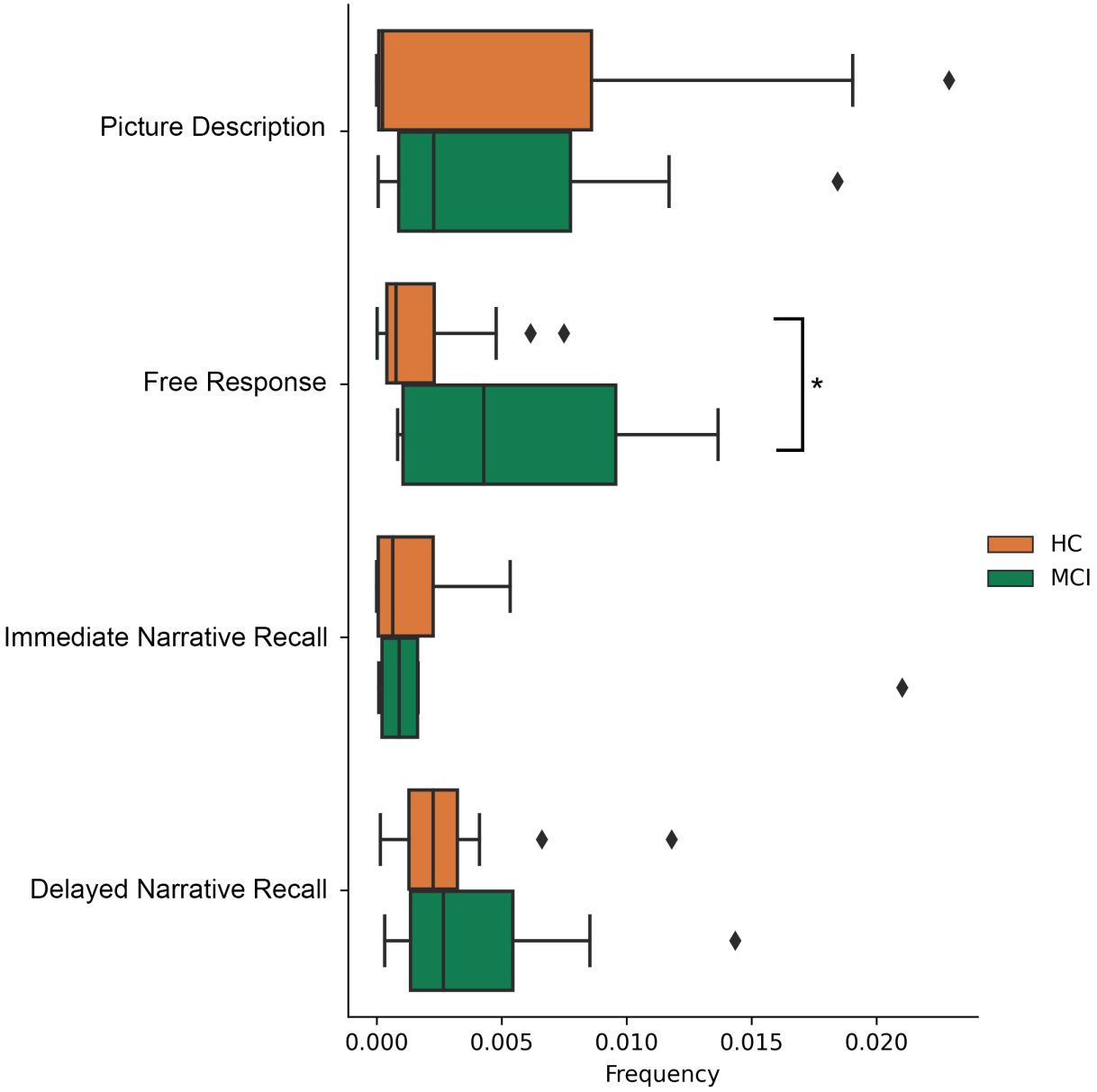

Figure 2. Task comparisons of post-"uh" disfluency word frequency rankings. This measure was used for comparison to highlight differences in disfluency and post-disfluency behavior between tasks. In both narrative recall tasks, all groups tended to be searching for comparatively uncommon words following "uh" disfluencies, while in the picture description task all groups searched for more common words. Only in the free response task were there differences between impairment status, with healthy controls (HC) searching only for more uncommon words and participants with mild cognitive impairment (MCI) producing words of more variable frequency following "uh" disfluencies.

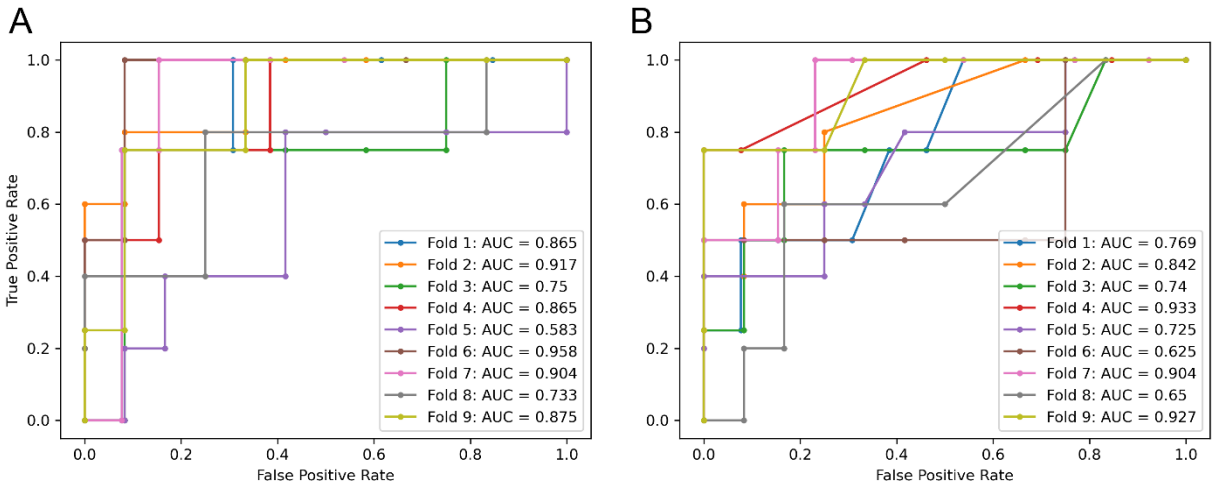

Figure 3. Areas under ROC curves for predictive models. (A) After performing feature selection on all available lexical features, 16 features were identified to best identify impairment status (HC vs MCI). Incorporating these into a repeated stratified 3-fold 3-repeat cross-validation procedure using *LightGBM* gradient boosted machines as the models generated a mean AUC of 0.828. AUCs for individual folds are depicted in multiple colors. (B) The best identified post-disfluency features (post-"uh" word frequency in free response and post-"um" latency in delayed narrative recall) were selected and used as sole features in a *LightGBM* gradient boosted machine, examined using repeated stratified 3-fold 3-repeat cross-validation. This parsimonious model performed similarly to the larger model with an AUC of 0.791. AUCs for individual folds are depicted in multiple colors.