HHS Public Access

Author manuscript

Gerontology. Author manuscript; available in PMC 2025 January 12.

Published in final edited form as:

Gerontology. 2024 ; 70(4): 429–438. doi:10.1159/000536250.

Harnessing Speech-Derived Digital Biomarkers to Detect and Quantify Cognitive Decline Severity in Older Adults

Gozde Cay[a], **Valeria A Pfeifer**[b], **Myeounggon Lee**[a], **Mohammad Dehghan Rouzi**[a], **Adonay S. Nunes**[c], **Nesreen El-Refaei**[a], **Anmol Salim Momin**[a], **MD Moin Uddin Atique**[a], **Matthias R. Mehl**[b], **Ashkan Vaziri**[c], **Bijan Najafi**[a]

[a]Interdisciplinary Consortium on Advanced Motion Performance (iCAMP), Michael E. DeBakey, Department of Surgery, Baylor College of Medicine, Houston, TX, USA

[b]Department of Psychology, University of Arizona, Tucson, AZ, USA

[c]BioSensics LLC, Newton, MA, USA

Abstract

Introduction: Current cognitive assessments suffer from floor/ceiling and practice effects, poor psychometric performance in mild cases, and repeated assessment effects. This study explores the use of digital speech analysis as an alternative tool for determining cognitive impairment. The study specifically focuses on identifying the digital speech biomarkers associated with cognitive impairment and its severity.

Methods: We recruited older adults with varying cognitive health. Their speech data, recorded via a wearable-microphone during the reading aloud of a standard passage, were processed to derive digital biomarkers such as timing, pitch, and loudness. Cohen's *D* effect size highlighted group differences, and correlations were drawn to the Montreal Cognitive Assessment (MoCA). A stepwise approach using a Random Forest model was implemented to distinguish cognitive states using speech data and predict MoCA scores based on highly correlated features.

Results: The study comprised 59 participants, with 36 demonstrating cognitive impairment and 23 serving as cognitively intact controls. Among all assessed parameters, similarity, as determined by Dynamic Time Warping (DTW), exhibited the most substantial positive correlation (rho=0.529, $p<0.001$), while timing parameters, specifically the ratio of extra words, revealed the strongest negative correlation (rho=−0.441, $p<0.001$) with MoCA scores. Optimal discriminative performance was achieved with a combination of four speech parameters: total pause time, speech

Corresponding Author: Full name: Bijan Najafi, Department: Interdisciplinary Consortium on Advanced Motion Performance (iCAMP), Michael E., DeBakey Department of Surgery, Institute/University/Hospital: Baylor College of Medicine, Street Name & Number: 7200 Cambridge St. Suite B01.529, City, State, Postal code, Country: Houston, TX, USA, bijan.najafi@bcm.edu.

Author Contributions

B.N conceived of the presented idea. N.R. and A.M. recruited the participants and collected the data. A.S.N. developed machine learning models for automatic speech metric detection with the supervision of A.V. G.C., M.A. and M.L. analyzed the speech data. M.D.R designed machine learning model and visualized machine learning results. B.N. supervised the project. V.P. and M.M. provided the background information and helped evaluate methods and results. All authors discussed the results and contributed to the final manuscript.

Study approval statement: This study protocol was reviewed and approved by Baylor College of Medicine Institutional Review Board, approval numbers H-43917, H-41717, and H-43413.

Consent to participate statement: A separate written informed consent was obtained from participants to participate in the study.

to pause ratio, similarity via DTW, and ratio of extra words. Precision and balanced accuracy scores were found to be 84.3 ± 1.5% and 75.0 ± 1.4%, respectively.

Discussion: Our research proposes that reading-derived speech data facilitates the differentiation between cognitively impaired individuals and cognitively intact, age-matched older adults. Specifically, parameters based on timing and similarity within speech data provide an effective gauge of cognitive impairment severity. These results suggest speech analysis as a viable digital biomarker for early detection and monitoring of cognitive impairment, offering novel approaches in dementia care.

Keywords

Cognition; cognitive decline; speech; wearables; speech assessment

Introduction

The increase in the aging population has led to the rise of the “dementia epidemic” affecting 55 million people worldwide [1]. This number is expected to grow to 135.5 million by 2050 [2]. Alzheimer’s disease (AD) is a degenerative brain disease and the most common form of dementia [3]. Despite numerous clinical trials for drug development conducted in the last 15 years, the failure rate of these trials is 99.6% [4]. The lack of robust tools for measuring and monitoring cognitive ability and function is considered a major factor contributing to the high failure rate. The currently available cognitive assessments have limited ecological validity [5], suffer from significant floor/ceiling effects, and remain prone to repeated assessment effects, leading to poor psychometric performance especially in mild cases [6]. Additionally, most of the current assessment tools do not represent an objective, reproducible assay for AD pathology and disease progression. This highlights the need for more effective and accurate methods of measuring cognitive decline in older adults.

Language functioning can provide valuable insights into cognition and behavior, serving as a window into an individual’s cognitive functioning [7–12]. Speech and language offer an opportunity to use ecologically valid methods to assess key symptoms related to cognitive function, a unique advantage in the assessment and monitoring of cognitive decline. Previous studies suggest that speech-based digital biomarkers can serve as reliable indices of AD severity and can be predictors of cognitive decline years before clinical diagnosis of AD [13–17]. Prior works in clinical trials also suggest that speech biomarkers are more sensitive to change than conventional cognitive screening methods such as the Mini-Mental State Examination (MMSE) [18, 19]. Additionally, a recent study suggests that the digital biomarkers of speech associated with cognitive changes over time are independent of practice or learning effects [19].

Despite the established value of speech assessment in identifying cognitive decline and assessing its severity, no practical solution currently exists for remote collection of corresponding digital speech biomarkers. This gap presents a pressing issue, particularly for individuals experiencing socioeconomic deprivation or those residing in remote areas who often have restricted access to traditional cognitive screening tools [7]. This inequity, be it

social, health-related, or environmental, adversely affects timely access to care, exacerbating health disparities and hindering effective management of dementia symptoms.

The growing prevalence of cognitive impairments, like dementia, coupled with the current limitations in diagnostic approaches, underscores the urgent need for a robust Remote Patient Monitoring (RPM) system. Such a system would be instrumental in the early identification of cognitive decline, enabling ongoing monitoring of its progression [11]. In this context, the development of a wearable device, functioning as an RPM tool, could significantly bridge this gap. By gathering speech-derived digital biomarkers, such device could not only discern cognitive impairment but also gauge its severity. The potential impact of such a tool is multifaceted. By offering remote cognitive monitoring, the barriers of distance and socioeconomic disparities could be mitigated, improving healthcare access for underserved populations. Furthermore, it would allow for a more continuous, real-time assessment of cognitive function, rather than relying on infrequent clinical visits. This would result in a more accurate understanding of the progression of cognitive impairment, facilitating more personalized and timely interventions.

Speech data tap into an aspect of cognition that is routinely engaged in daily life and is highly sensitive to cognitive changes. Thus, speech-derived digital biomarkers could provide a more immediate and real-world relevant measure of cognitive function. This could result in a more nuanced understanding of cognitive decline progression and its impact on daily living, leading to better patient-centered care and management strategies. Therefore, while more research in this area is needed, the potential implications are profound, offering new avenues in dementia care, facilitating early detection, and enabling more personalized and effective care and management strategies.

As an initial step towards designing such a RPM system to detect and track cognitive decline using speech data, we evaluated whether clinically meaningful speech-derived digital metrics could be extracted from data collected using an off-the-shelf wearable microphone, without imposing strict control over environmental conditions such as microphone distance, but while recording at a clinical facility. Additionally, we developed an algorithm to automate the extraction of speech data, eliminating the need to manually monitor the speech recording to determine the start and end points of a standard speech task. The aim was to identify and quantify cognitive impairment with these automated and adaptable processes, laying the groundwork for a practical, user-friendly RPM system in the future.

Our hypothesis was that speech characteristics, derived from reading a standardized passage aloud and recorded through a wearable microphone in the lab, could effectively differentiate individuals with cognitive impairment from cognitively healthy, age-matched controls. Additionally, we posited that a composite of these speech-derived metrics could serve as a robust tool for gauging the severity of cognitive impairment.

Materials and Methods

Participant Recruitment

This study presents a secondary analysis of data from three previous cohorts, with participants enrolled from Baylor College of Medicine's Pulmonary, Vascular Surgery, Intensive Care Unit, Primary Care, and Psychiatry departments. We included older adults (50 years and over) with cognitive impairment and age-matched cognitively healthy counterparts. Inclusion criteria mandated available speech data from participants reading the standard Rainbow Passage aloud [20], along with relevant clinical data for cognitive function and impairment assessment. Speech data were collected using a wearable microphone in a supervised laboratory setting. We identified cognitive impairment via clinical diagnosis from electronic health record, if absent, a Montreal Cognitive Assessment (MoCA) score below 26 [21]. The only exclusion criterion was inability to fluently read and speak in English. The Baylor College of Medicine Institutional Review Board approved the study protocol (protocol numbers H-43917, H-41717, and H-43413).

Test protocols

Participants gave a separate written informed consent, then completed a demographic background questionnaire and a number of additional questionnaires: 1) MoCA to assess cognitive function; 2) Falls Efficacy Scale (FES-I) to assess risk of falling [22]; 3) Center for Epidemiologic Studies Depression (CES-D) scale to determine the level of depressive symptoms [23]; and 4) Beck Anxiety Inventory (BAI) to evaluate the level of anxiety symptoms [24]. Last, participants completed the Speech Assessment described below, and finally, received debriefing. For the Speech Assessment, participants were asked to read out loud the rainbow passage [20] (full text in Appendix A) from a laminated card with the text printed on it at their own pace. If they did not finish the reading task within 60 seconds, the task was ended by the experimenter. Their speech was recorded using a wearable microphone (Olympus Linear PCM Recorder LS-P4, OM Digital Solutions, Shinjuku, Tokyo, Japan) attached to their collar. The data were recorded at 32 kHz and files were exported from the recorder as .wav file via Audacity audio editor software [25].

Extraction of Speech Parameters

Firstly, the audio files were denoised using an open-source audio editor software (Ocenaudio, Federal University of Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brazil) [26] to filter the background noise (See Supplementary Figure 1). Then the denoised speech data were analyzed with an automated software solution to extract features from speech assessments (BioDigit Speech, BioSensics LLC, Newton, MA USA.) The software automatically detects the relevant portion of audio involving the Rainbow Passage, and eliminates the remaining parts, i.e., where participants talk to the experimenter. The duration is only for the actual passage reading. BioDigit Speech uses automated speech recognition (ASR) to transcribe the speech and provide timestamps for each segment. In addition, silent segments are detected by estimating the audio intensity and thresholding silent periods that are 25 dB lower than the maximum intensity.

BioDigit Speech extracted several features from the passage reading (see Supplementary Table 1). These parameters are grouped in three categories as follows: Timing features includes the following features: the total pause time, total voiced time, and total signal time, measured as separate features. The articulatory rate, indicating the number of syllables articulated per second, was obtained by dividing the number of uttered syllables by the total voiced time. The mean pause length was calculated to determine whether the speaker tends to take longer or shorter pauses, and the total number of pauses was counted. The speech to pause ratio was also calculated by normalizing the voiced time by the pause time, irrespective of the total signal duration, and providing the proportion of speech in relation to pauses or silence. Three acoustic features were also extracted: average loudness, mean pitch (fundamental frequency), and pitch standard deviation (SD). These features are relevant because decreased pulmonary capacity can impact loudness, while neuromotor difficulties in regulating the vocal fold can affect pitch and pitch variability over time.

The transcription of the reading was compared to the actual passage's word content to extract the similarity features. Two features were calculated: the ratio of extra words and the ratio of missing words. The former indicates the number of words uttered by the participant during the Rainbow Passage reading that were not originally in the passage, divided by the total number of words read. Extra words could be due to subjects repeating words multiple times or misreading them. The latter feature indicates the number of words the subject missed in the passage compared to the total number of words read. Dynamic time warping (DTW) was utilized to compare the transcribed reading and the original passage. The DTW approach quantifies features in an automated and objective way, by looking at the distance between two speech signals. Two types of DTW were extracted: similarity DTW and intelligibility DTW. The former represents the 1/(1 + DTW distance) between the original passage and the transcribed reading, with higher values indicating greater similarity between the two signals. The latter measures the similarity between the transcription generated by an ASR model of medium size and a model of small size. The rationale behind using models of different complexity is that smaller models may struggle to accurately transcribe unclear speech, leading to lower Intelligibility DTW values. Instead of encoding words, letters were assigned numerical codes as it was suggested that this method better captures speech alterations [27, 28].

Statistical Analysis

A Shapiro-Wilk test was used to assess the normality of the data ($p > 0.05$). A chi-square test was used to compare categorical variables between the cognitively impaired and intact groups.

An Independent t-test was used in order to compare groups regarding demographics, clinical information, and speech features. If the assumption of normal distribution was not satisfied, Mann-Whitney U tests were used. Cohen's d was used to calculate the effect size. In addition, p-values for 14 speech features were adjusted according to the false discovery rate (FDR) [29].

Additionally, Pearson's product-moment correlation analysis was performed in order to investigate the relationship between all of the speech features and the MoCA score. When

the variable did not conform to the normal distribution, Spearman's rank correlation analysis was used. The statistical analyses were conducted using SPSS 28.0 (IBM, Chicago, IL, USA) and 0.05 was set as the level of statistical significance.

Random Forest Model Approach for Feature Ranking and Training

Artificial intelligence and machine learning techniques are increasingly influencing diverse medical fields, from subtle applications in conditions like chronic limb-threatening ischemia (CLTI) [30] to more pronounced impacts in areas like speech analysis [31]. Among these AI techniques, Random Forest modeling approach uses ensemble learning to provide solutions to complex problems by combining multiple classifiers [32, 33]. Notably, it has been successfully applied for the determination of feature importance, as evidenced in our prior work where we employed it for ranking features in the detection of physical aggression in ADHD children [34]. We used a Random Forest classifier to investigate whether using exclusively speech features can provide accurate predictions of a patient's cognitive status (intact versus impaired). The feature importance is computed based on the average decrease in impurity (usually Gini impurity) that each feature causes when used in trees of the forest, giving a measure of the contribution of each feature to the predictive power of the model. The feature ranking process involved two key steps. In the first step, all speech features that demonstrated a statistically significant correlation with the MoCA score were selected for the machine learning analysis. They were then ranked by training the Random Forests machine learning model.

The second step involved the utilization of a Random Forest Regression model to predict MoCA scores. This was facilitated by a feature elimination technique [35] which added features sequentially, commencing with the most significant feature. During each incremental step, we added one feature to the dataset, then trained the model to assess the impact of that feature on the model's predictive capacity. The scores predicted by this model were used to classify the participants into either cognitively impaired or intact cohorts. Additionally, the model estimates machine learning performance metrics including accuracy, precision, recall, F1 score, and balanced accuracy. Owing to the unbalanced distribution of our dataset, we opted for the balanced accuracy metric [36]. This metric is particularly advantageous in scenarios where the standard accuracy metric could be inflated due to precise predictions of the majority class; in contrast, balanced accuracy would decrease under such circumstances, thereby providing a more realistic representation of the model's predictive ability [37].

The training and testing sample sizes of the machine learning model were 80% and 20%, respectively. The Random Forest model was trained with 500 trees with a balanced subsample for class weights to improve performance [38, 39]. This approach was executed with 100 bootstraps, ensuring that all observations are part of the validation sub-sample and facilitating the calculation of means and standard deviations to quantify uncertainties [40–42]. The machine learning processes were conducted using Python version 3.10 (Python Software Foundation, Fredericksburg, VA, USA), leveraging numpy and pandas libraries for data handling, and scikit-learn for machine learning functionalities.

Results

Participants characteristics

Our study comprised fifty-nine older adults (N = 59) who met the designated inclusion and exclusion criteria. Based on clinical diagnosis and MoCA score, 36 participants were classified as cognitively impaired (Mean age = 67 ± 9, sex = 69.44% female, Mean BMI = 29 ± 6, Mean MoCA score = 21 ± 4, and the remaining 23 were grouped as age-matched cognitively intact individuals (Mean age = 69 ± 6, sex = 60.87% female, Mean BMI = 29 ± 5.84, Mean MoCA score = 26 ± 3). The demographic profile, patient-reported outcomes, and comorbidity rates demonstrated no significant differences between the groups, except in MoCA scores ($p < 0.001$, $d = 1.33$; refer to Table 1 for details)

Speech Analysis

The cognitive impairment group exhibited greater trends for total pause time ($p=0.029$, $d=0.593$) and the speech to pause ratio ($p=0.017$, $d=0.656$) (See Supplementary Table 2). However, if we considered the adjusted p-value according to the FDR, there was no significant difference between groups in the total pause time (p-adjustment= 0.20) and speech to pause ratio (p-adjustment=0.234). There were no significant differences in pitch, loudness, and similarity features.

The total signal time (rho = −0.316, p=0.015), number of pause (rho = −0.343, p=0.008), total pause time (rho = −0.406, p=0.001), ratio extra words (rho = −0.441, p<0.001), and ratio missing words (rho = −0.393, p=0.002) were negatively correlated with the MoCA score; whereas the speech to pause ratio (rho = 0.381, p=0.003), similarity DTW (rho = 0.529, p<0.001), and intelligibility DTW (rho = 0.333, p=0.010) were positively correlated with the MoCA score (Figure 1).

Optimal Feature Selection and Evaluation

Eight speech features which were significantly correlated with MoCA (shown in Figure 1) were selected to train our machine learning model. Figure 2(A) shows the importance ranking of these features examined using the Random Forest machine learning model. The greater the percentage value, the greater the contribution to the model's output made by the feature. Similarity DTW ranks highest, while the ratio of missing words ranks lowest.

Figure 2(B) presents the model validation results as a function of the number of ranked features. The model with all eight features achieved an accuracy of 78.3 ± 1.1%, precision of 87.7 ± 1.2%, recall of 83.1 ± 1.3%, f1 score of 84.4 ± 0.9%, and balanced accuracy of 74.8 ± 1.4%. The model achieved highest precision (88.1 ± 1.2%) and balanced accuracy (76.3 ± 1.3%) by incorporating four features, namely similarity DTW, speech to pause ratio, intelligibility DTW, and total pause time, while maintaining a satisfactory range (0.7 to 0.9) for accuracy (77.5 ± 1.0%), recall (81.0 ± 1.1%), and f1 score (83.5 ± 0.9%). Using four features in the model resulted in a mean absolute error of 3.0 ± 0.1 and an explained variance of 0.2 ± 0.1 in the regression model.

Discussion

In this study, we investigated whether voice-driven digital metrics assessed during passage reading obtained with wearable microphones and a novel processing pipeline, including the use of DTW features, are associated with cognitive decline severity in older adults, laying the groundwork for a practical, user-friendly RPM system. Our results support the hypothesis that analyzing temporal speech features during reading aloud of a standard passage, captured by the wearable microphone, can help differentiate individuals with cognitive impairment and assess the severity of impairment. Specifically, among the 14 speech digital metrics assessed in this study, both total pause time and speech to pause ratio achieved a medium effect size range in distinguishing between groups with and without cognitive impairment. Furthermore, univariate analysis revealed significant correlations between 8 digital speech metrics and MoCA scores, allowing for a direct prediction of MoCA scores from speech metrics. Lastly, we developed a machine learning model to predict cognitive performance using these speech-driven digital biomarkers. The results suggest that by using four speech digital metrics, including similarity DTW, speech to pause ratio, intelligibility DTW, and total pause time, the model can classify participants into ‘cognitively impaired’ and ‘cognitively intact’ categories with a precision of 88.1% and a balanced accuracy of 76.3%. The relatively good performance achieved here, even with a small sample size, might be attributed to our feature selection process. We emphasize the robustness of our Random Forest model and how it effectively ranked and chose the most significant features that could have contributed to this enhanced performance.

Our results showed that total pause time and speech to pause ratio are the features that differ significantly between cognitively impaired and intact groups. This is in agreement with prior studies analyzing free speech, not read-aloud passages. For example, Tóth and colleagues extracted several acoustic parameters from the speech of patients with mild cognitive impairment to find out which features correlated most with severity [43]. They found that, amongst others, speech rate, as well as number of length of pauses differed significantly between the healthy controls and the cognitive intact group, suggesting they are ‘acoustic biomarker’ of cognitive decline.

Additionally, we did not find a correlation between articulatory rate, total voiced time, and MoCA score, suggesting that it is not slowness of speech in itself that is related to cognitive impairment, but likely cognitive processes associated with the speech planning or reading. Future work could tease apart these two possibilities. Our results also suggested that physical properties of the recorded speech, such as loudness, are not related to cognitive impairment, which demonstrates that wearable technology can be readily used for this type of cognitive assessment since the location and quality of the microphone in our study did not affect the assessment of cognition. Moreover, our findings suggested that the similarity parameter, which represents the difference between the original passage and the produced reading, is related to cognitive impairment, suggesting that features inherent to passage reading, not the recording, matter. Specifically, individuals with cognitive impairment show a lower level of similarity between produced speech and the passage, suggesting once more that our method captures cognition, not physical features of the speech behavior. While we cannot differentiate where in the reading process difficulties might occur (e.g., accurately

decoding the visual word form while reading, accessing the concepts stored in the mental lexicon, or mapping the concepts onto their phonological representation), the similarity parameter has the potential to be a reliable and innovative marker for evaluating cognitive functioning and can be readily captured.

This study establishes a preliminary foundation for the creation of a RPM system that could acquire speech-derived digital biomarkers, indicative of cognitive impairment, via telemedicine or wearable technologies. Through the development of automated processes, it's possible to isolate speech from environmental noise and extract meaningful digital biomarkers without the need for human intervention or direct speech listening. Upon noise filtering, our utilized BioDigit Speech software aids in automatically recognizing and exporting the designated read-aloud passages from audio files, neglecting any irrelevant speech. This automation streamlines the process, eliminating the necessity for human annotators and potentially enabling asynchronous, remote cognitive testing—an advantageous attribute for diagnosing individuals in geographically remote areas. However, while promising, the proposed remote speech assessment approach necessitates further validation through additional research. Interestingly, our findings suggest that it isn't voice-based characteristics, but rather timing and similarity metrics within speech data, that effectively differentiate between cognitively impaired individuals and age-matched cognitively intact controls, further underlining the necessity of passage reading, not free speech, for measuring cognitive decline. Moreover, since the every participant is reading the same content, passage reading ensures consistency across all participants and somparison becomes straightforward. Additionally, these metrics show potential for quantifying the severity of cognitive impairment, underscoring their potential value in cognitive health assessments. Together, these results suggest that a basic microphone could suffice in creating a cost-effective wearable system for remote cognitive function monitoring, and that an automated processing pipeline can yield reliable results in seconds. Developing such a system could revolutionize cognitive health management by facilitating continuous, remote cognitive assessment, making cognitive health monitoring more accessible, efficient, and equitable. An additional advantage is the potential for improved data privacy; rather than recording comprehensive speech data, only de-identified speech characteristics—such as timing and similarity features—are recorded, potentially assuaging privacy concerns.

In this study, we utilized the Random Forest algorithm rooted in its alignment with the clinical orientation and the objective of our study, which aims to construct a reliable and practical Remote Patient Monitoring (RPM) system [44, 45]. Random Forest presents as a well-established, interpretable machine learning tool that offers an efficient feature selection mechanism, aiding in the optimal extraction of significant predictors of cognitive decline from speech data [46, 47]. To ensure the robustness of our classification model and to provide a comprehensive perspective, we extended our analysis to include SVM (Support Vector Machine) [48] and Gaussian Naive Bayes classifiers. Using the balanced accuracy metric, a choice made due to our dataset's unbalanced distribution, we found the Random Forest model had a balanced accuracy of 74.8 ± 1.4%. In comparison, the SVM model registered a balanced accuracy of 72.9 ± 1.2%, and the Gaussian Naive Bayes model showed a balanced accuracy of 72.4 ± 1.3%. These findings underscore the comparative effectiveness of the Random Forest model while also illustrating the relative performance

of other classifiers. It is worth noting that deep learning models, renowned for their performance across various domains, typically require large datasets to achieve their full potential and to minimize the risk of overfitting. Given that our dataset consists of only 59 subjects, the application of deep learning models might be less optimal [49]. Therefore, in contexts with limited data, traditional machine learning models, such as those utilized in our study, often provide more consistent and reliable results.

A few limitations to generalizability arise. First, our sample primarily consists of white individuals, and our sample size of 59 participants did not allow us to further break down the sample. A larger sample size could help shed light on whether race or sex differences might moderate effects of cognition on speech. Further, our participants were sampled as a secondary analysis of existing data. Our inclusion criteria were determined according to the parent studies which may have led to the exclusion of participants who were eligible for speech assessments. Further studies which focus on speech assessment as the primary objective could further verify that these exclusions did not affect our findings.

A second limitation arises from the use of passage reading. While this approach allows us to develop the similarity metric that ultimately contributed greatly to the prediction of cognitive impairment, reading ability might be confounded with a number of different variables, such as education, first language, or vision impairments of our participants. In addition, it might not capture the spontaneous speech patterns that occur in day-to-day conversation. However, we believe that as an initial step, our approach seems promising. Future research endeavors should aim to transpose our approach from the clinical setting into participants' homes. The current study had participants read aloud in a controlled, supervised environment within an outpatient clinical setting. A key objective for subsequent work would be to collect audio data from participants who independently read the Rainbow Passage and recorded themselves using a tablet or other personal device. We envision that the code we utilized for file processing and speech feature extraction could be readily integrated into a tablet or phone application, which would then facilitate remote cognitive assessments in unsupervised settings, thus enhancing the accessibility, convenience, and scalability of cognitive health monitoring.

Conclusion

In summary, our study reveals that speech data, extractable via a commercially available wearable microphone and particularly emphasizing timing and similarity-based measures, serves a dual purpose. It not only differentiates individuals with cognitive impairment from age-matched cognitively intact older adults, but also enables quantification of cognitive impairment severity. This suggests that such speech-derived metrics offer a promising approach for both identifying and monitoring the progression of cognitive decline. The emphasis on timing and similarity-based aspects in speech analysis in cognitive impairment assessment brings several benefits. These include the use of cost-effective microphones, relaxed constraints on environmental conditions like noise levels, microphone distance, and speech loudness standardization, thereby facilitating its deployment for remote patient monitoring via telemedicine, smart home infrastructure (based on Internet of Things), or wearable technology. Furthermore, our research underscores the feasibility of automated

processes for determining the beginning and end of speech data and extracting digital parameters of interest, thus reducing the need for human intervention or direct listening to the speech data. This method potentially enhances data privacy, as the system could primarily record speech digital biomarkers (like timing and similarity metrics) rather than the speech data itself, effectively producing de-identified metrics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors thank Kala Pham for organizing the data, and Maria Noun for communication with IRB.

Funding Sources

Research reported in this publication was supported by the National Institute for Aging of the National Institutes of Health under Award Number R44AG061951 and U19AG065169. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest Statement

A.S.N. and A.V. are associated with BioSensics LLC, his role being limited to providing technical support and the development of machine learning models for automatic speech metric detection. M.A is now with Abbott. However his contributions to this study were limited to when he was postdoctoral associate with Baylor College of Medicine. B.N. served as a consultant for BioSensics on projects unrelated to the scope of this project. Although he did not participate in patient recruitment or data analysis, he made significant contributions to the study design, acquisition of funding, interpretation of results, and manuscript revision. The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability Statement

The raw data supporting the conclusions of this study are not publicly accessible due to privacy concerns, as speech data can be used to identify individuals. However, processed de-identified data that do not compromise participant anonymity are available upon formal request to the senior author.

References

1. Kelaiditi E, Cesari M, Canevelli M, van Kan GA, Ousset PJ, Gillette-Guyonnet S, et al. Cognitive frailty: rational and definition from an (I.A.N.A./I.A.G.G.) international consensus group. J Nutr Health Aging. 2013 Sep;17(9):726–34. [PubMed: 24154642]
2. Mental health of older adults. World Health Organization; 2017.
3. Health Equity.
4. Cummings J, Lee G, Nahed P, Kambar M, Zhong K, Fonseca J, et al. Alzheimer's disease drug development pipeline: 2022. Alzheimers Dement (N Y). 2022;8(1):e12295. [PubMed: 35516416]
5. Spooner DM, Pachana NA. Ecological validity in neuropsychological assessment: a case for greater consideration in research with neurologically intact populations. Arch Clin Neuropsychol. 2006 May;21(4):327–37. [PubMed: 16769198]
6. Kueper JK, Speechley M, Montero-Odasso M. The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review. J Alzheimers Dis. 2018;63(2):423–44. [PubMed: 29660938]

7. Rektorova I, Mekyska J, Janousova E, Kostalova M, Eliasova I, Mrackova M, et al. Speech prosody impairment predicts cognitive decline in Parkinson's disease. Parkinsonism Relat Disord. 2016 Aug;29:90–5. [PubMed: 27237105]
8. Beltrami D, Gagliardi G, Rossini Favretti R, Ghidoni E, Tamburini F, Calza L. Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline? Front Aging Neurosci. 2018;10:369. [PubMed: 30483116]
9. Ambrosini E, Caielli M, Milis M, Loizou C, Azzolino D, Damanti S, et al. Automatic speech analysis to early detect functional cognitive decline in elderly population. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)2019. p. 212–16.
10. Konig A, Mallick E, Troger J, Linz N, Zeghari R, Manera V, et al. Measuring neuropsychiatric symptoms in patients with early cognitive decline using speech analysis. Eur Psychiatry. 2021 Oct 13;64(1):e64. [PubMed: 34641989]
11. Pan Y, Nallanthighal VS, Blackburn D, Christensen H, Härmä A. Multi-Task Estimation of Age and Cognitive Decline from Speech. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)2021. p. 7258–62.
12. Wang HL, Tang R, Ren RJ, Dammer EB, Guo QH, Peng GP, et al. Speech silence character as a diagnostic biomarker of early cognitive decline and its functional mechanism: a multicenter cross-sectional cohort study. BMC Med. 2022 Nov 7;20(1):380. [PubMed: 36336678]
13. Snowdon DA. Aging and Alzheimer's disease: lessons from the Nun Study. Gerontologist. 1997 Apr;37(2):150–6. [PubMed: 9127971]
14. Garrard P, Maloney LM, Hodges JR, Patterson K. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain. 2004;128(2):250–60. [PubMed: 15574466]
15. van Velzen M, Garrard P. From hindsight to insight – retrospective analysis of language written by a renowned Alzheimer's patient. Interdisciplinary Science Reviews. 2008 2008/12/01;33(4):278–86.
16. Le X, Lancashire I, Hirst G, Jokel R. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Literary and Linguistic Computing. 2011;26(4):435–61.
17. Vigo I, Coelho L, Reis S. Speech- and Language-Based Classification of Alzheimer’s Disease: A Systematic Review. Bioengineering. 2022;9(1):27. [PubMed: 35049736]
18. Yancheva M, Fraser KC, Rudzicz F. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies2015. p. 134–39.
19. Simpson W, Kaufman LD, Detke M, Lynch C, Butler A, Dominy S. Utility of speech-based digital biomarkers for evaluating disease progression in clinical trials of Alzheimer's disease. Alzheimer's & Dementia. 2019;15(7):P1524.
20. Fairbanks G. Voice and articulation drillbook. Addison-Wesley Educational Publishers; 1960.
21. Nasreddine ZS, Phillips NA, Bedirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005 Apr;53(4):695–9. [PubMed: 15817019]
22. Yardley L, Beyer N, Hauer K, Kempen G, Piot-Ziegler C, Todd C. Development and initial validation of the Falls Efficacy Scale-International (FES-I). Age Ageing. 2005 Nov;34(6):614–9. [PubMed: 16267188]
23. Radloff LS. A self-report depression scale for research in the general population. Applied psychological measurement. 1977;1:385–401.
24. Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol. 1988 Dec;56(6):893–7. [PubMed: 3204199]
25. Audacity.
26. Ciesla R. Voice Acting: Hardware and Techniques. Sound and Music for Games: The Basics of Digital Audio for Video Games. Springer; 2022. p. 127–46.

27. Orozco-Arroyave JR, Vásquez-Correa JC, Vargas-Bonilla JF, Arora R, Dehak N, Nidadavolu PS, et al. NeuroSpeech: An open-source software for Parkinson's speech analysis. Digital Signal Processing. 2018;77:207–21.
28. Kang K, Nunes AS, Sharma M, Hall A, Mishra RK, Casado J, et al. Utilizing speech analysis to differentiate progressive supranuclear palsy from Parkinson's disease. Parkinsonism & Related Disorders. 2023:105835. [PubMed: 37678101]
29. Benjamini Y. Discovering the false discovery rate. Journal of the Royal Statistical Society: series B (statistical methodology). 2010;72(4):405–16.
30. Bagheri AB, Rouzi MD, Koohbanani NA, Mahoor MH, Finco M, Lee M, et al. Potential applications of artificial intelligence (AI) and machine learning (ML) on diagnosis, treatment, outcome prediction to address health care disparities of chronic limb-threatening ischemia (CLTI). Seminars in Vascular Surgery: Elsevier; 2023.
31. Deshpande G, Schuller B. An overview on audio, signal, speech, & language processing for COVID-19. arXiv preprint arXiv:200508579. 2020.
32. Pal M. Random forest classifier for remote sensing classification. International journal of remote sensing. 2005;26(1):217–22.
33. Rogers J, Gunn S. Identifying feature relevance using a random forest. Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop, SLSFS 2005, Bohinj, Slovenia, February 23-25, 2005, Revised Selected Papers: Springer; 2006. p. 173-84.
34. Park C, Rouzi MD, Atique MMU, Finco M, Mishra RK, Barba-Villalobos G, et al. Machine Learning-Based Aggression Detection in Children with ADHD Using Sensor-Based Physical Activity Monitoring. Sensors. 2023;23(10):4949. [PubMed: 37430862]
35. Lee H, Joseph B, Enriquez A, Najafi B. Toward Using a Smartwatch to Monitor Frailty in a Hospital Setting: Using a Single Wrist-Wearable Sensor to Assess Frailty in Bedbound Inpatients. Gerontology. 2018;64(4):389–400. [PubMed: 29176316]
36. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. 2010 20th international conference on pattern recognition: IEEE; 2010. p. 3121–24.
37. Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology: the Official Publication of the International Genetic Epidemiology Society. 2007;31(4):306–15.
38. Sage A. Random forest robustness, variable importance, and tree aggregation. 2018.
39. Wang Z, Jiang C, Ding Y, Lyu X, Liu Y. A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. Electronic Commerce Research and Applications. 2018;27:74–82.
40. Tibshirani RJ, Efron B. An introduction to the bootstrap. Monographs on statistics and applied probability. 1993;57(1).
41. Zhu W. Making bootstrap statistical inferences: A tutorial. Research quarterly for exercise and sport. 1997;68(1):44–55. [PubMed: 9094762]
42. Adelabu S, Mutanga O, Adam E. Testing the reliability and stability of the internal accuracy assessment of random forest for classifying tree defoliation levels using different validation methods. Geocarto International. 2015;30(7):810–21.
43. Tóth L, Hoffmann I, Gosztolya G, Vincze V, Szatlóczki G, Bánréti Z, et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Current Alzheimer Research. 2018;15(2):130–38. [PubMed: 29165085]
44. Sreejith S, Rahul S, Jisha R. A real time patient monitoring system for heart disease prediction using random forest algorithm. Advances in Signal Processing and Intelligent Recognition Systems: Proceedings of Second International Symposium on Signal Processing and Intelligent Recognition Systems (SIRS-2015) December 16-19, 2015, Trivandrum, India: Springer; 2016. p. 485–500.
45. Kaur P, Kumar R, Kumar M. A healthcare monitoring system using random forest and internet of things (IoT). Multimedia Tools and Applications. 2019;78:19905–16.

46. Zheng L, Li Q, Ban H, Liu S. Speech emotion recognition based on convolution neural network combined with random forest. 2018 Chinese control and decision conference (CCDC): IEEE; 2018. p. 4143–47.

47. Chen L, Wu M, Pedrycz W, Hirota K, Chen L, Wu M, et al. Two-Layer Fuzzy Multiple Random Forest for Speech Emotion Recognition. Emotion Recognition and Understanding for Emotional Human-Robot Interaction Systems. 2021:77–89.

48. Steinwart I, Christmann A. Support vector machines. Springer Science & Business Media; 2008.

49. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems. 2022;35:507–20.

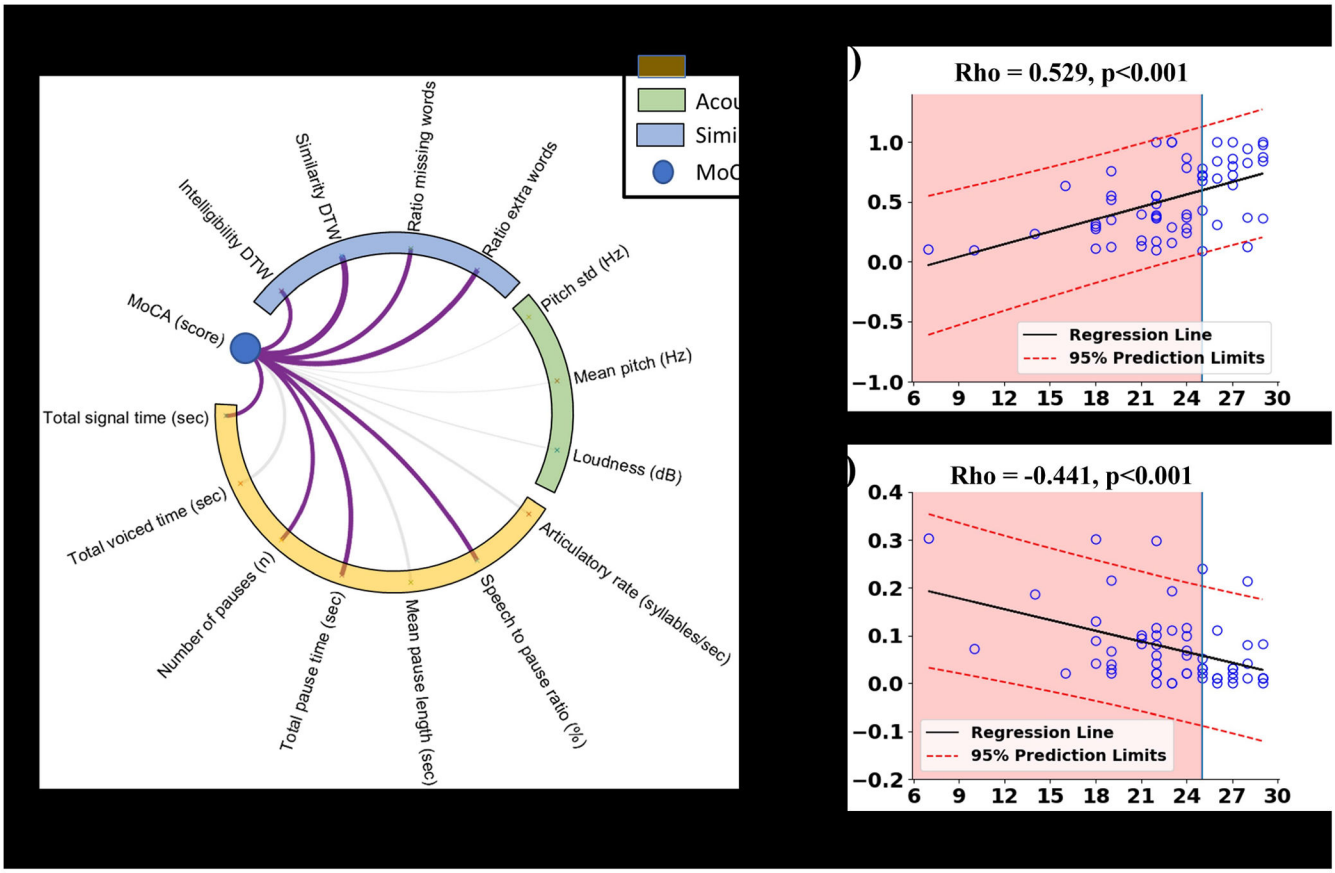

Figure 1:

Correlation results between speech features and MoCA score: (a) is a correlation between MoCA and speech assessment parameters, purple lines indicate significant correlations and other lines are not significant; (b) and (c) are scatter plots exhibited

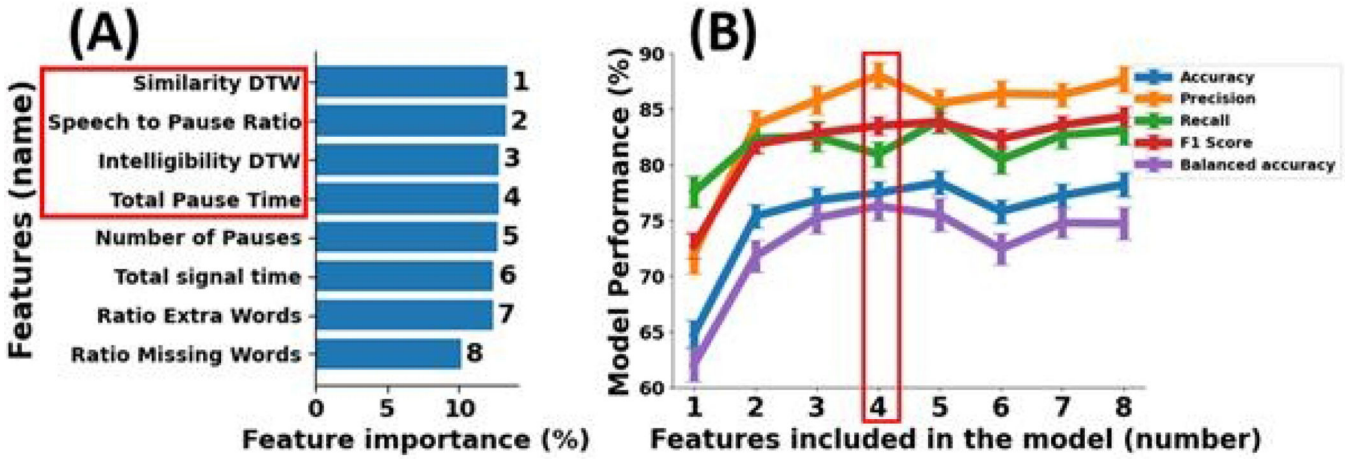

Figure 2:

(A) shows the ranking of eight features based on their significance in identifying cognitive impaired subjects from cognitive intact subjects, as determined by the Random Forest machine learning model. Furthermore, (B) illustrates the effectiveness of the trained machine learning model in distinguishing between two groups according to the accuracy, precision, recall, f1 score, and balanced accuracy of the model.

Table 1.

Demographic and clinical information

Variable	Cognitive impairment (n=36)	Cognitive intact (n=23)	P-value	Effect-size: Cohen's d
Demographic information				
Age (years)	67 (9)	69 (6)	0.629	0.13
Sex: Female, n (%)	25 (69.44%)	14 (60.87%)	0.497	0.18
BMI (kg/m^2)	29 (6)	29 (6)	0.822	0.06
Ethnicity: Hispanic, n (%)	8 (22.22%)	4 (17.39%)	0.653	0.12
Race, n (%)				
White	16 (44.44%)	12 (52.17%)		
Black	25 (50.00%)	10 (43.48%)	0.843	0.15
Asian	2 (5.56%)	1 (4.35%)		
Education level, n (%)				
Less than high school	1 (2.78%)	2 (8.69%)		
High school degree	16 (44.44%)	8 (34.78%)	0.653	0.42
Associate degree	10 (27.78%)	7 (30.43%)		
Above Bachelor's degree	9 (25.00%)	6 (26.09%)		
Patient-reported outcomes				
Cognitive function: MoCA (score)	21 (4)	26 (3)	**<0.001**	1.33
Risk of falls: FES-I (score)	24 (8)	26 (8)	0.204	0.34
Depression: CES-D (score)	14 (9)	16 (5)	0.326	0.26
Anxiety level: BAI (score)	9 (12)	10 (12)	0.730	0.09
Comorbidity				
High blood pressure, n (%)	21 (58.33%)	10 (43.48%)	0.891	0.04
Heart disease, n (%)	8 (22.22%)	4 (17.39%)	0.653	0.12
Stroke, n (%)	3 (8.33%)	1 (4.35%)	0.553	0.16
Depression, n (%)	7 (19.44%)	6 (26.09%)	0.548	0.16
Problem sleeping, n (%)	13 (36.11%)	11 (47.83%)	0.372	0.23
Cancer, n (%)	5 (13.89%)	5 (21.74%)	0.433	0.21
Had falling last year, n (%)	8 (22.22%)	10 (43.78%)	0.084	0.46

Mean and (std). BAI: Beck anxiety inventory. BMI: Body mass index. CES-D: Center for Epidemiologic Studies Depression Scale. FES-I: The Falls Efficacy Scale International. MoCA: The Montreal Cognitive Assessment. Mann-Whitney U-test: Comparisons between Cognitive impairment and cognitive intact groups; # was performed using an Independent t-test; Pearson's Chi-square test: Comparisons between two groups in categorical variables; significant level is a $p<0.05$. Effect-size was calculated by Cohen's d.