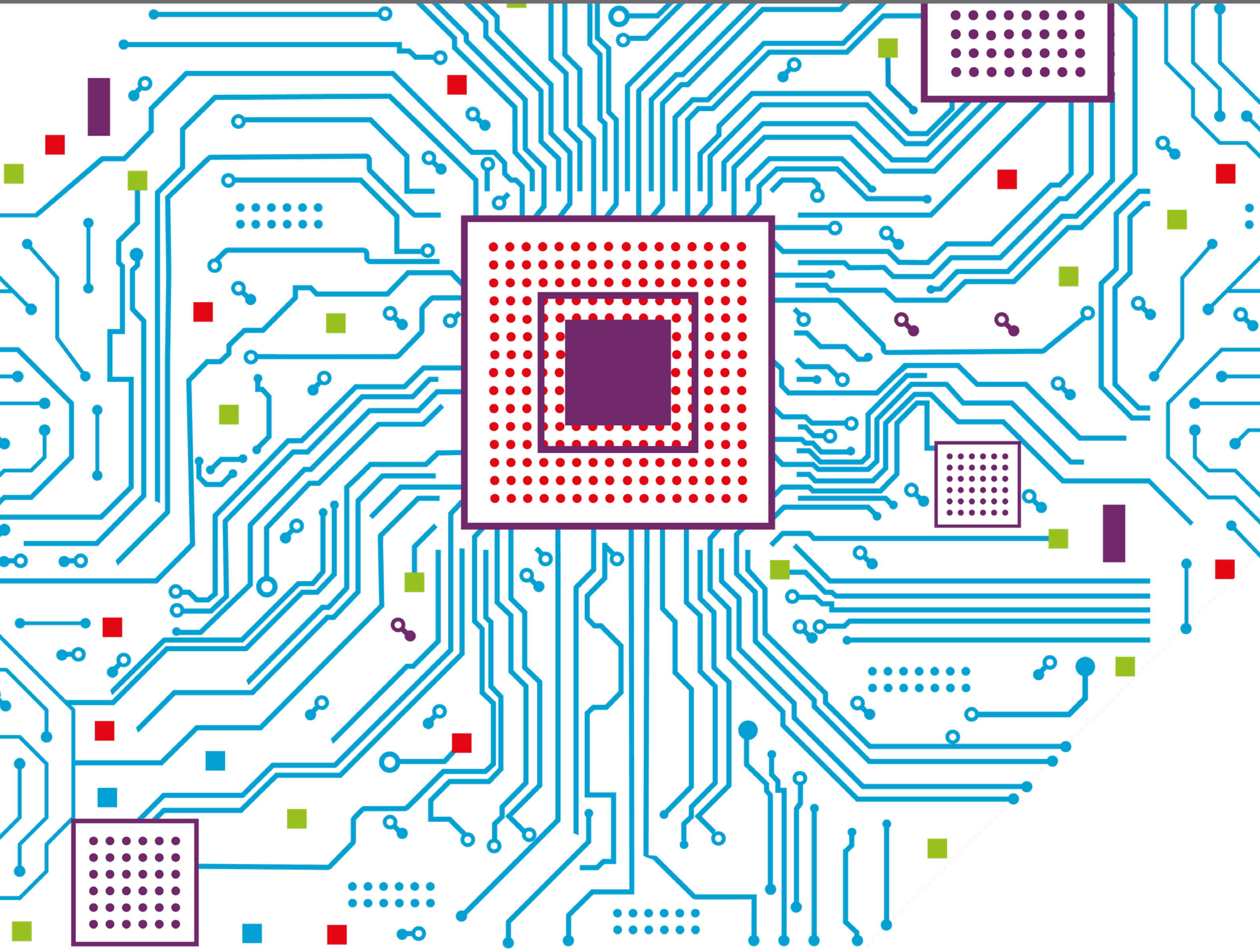
ALZHEIMER'S DEMENTIA RECOGNITION THROUGH SPONTANEOUS SPEECH

EDITED BY: Saturnino Luz, Fasih Haider, Davida Fromm and Brian MacWhinney

PUBLISHED IN: Frontiers in Computer Science, Frontiers in Psychology and Frontiers in Aging Neuroscience



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-854-2
DOI 10.3389/978-2-88971-854-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ALZHEIMER'S DEMENTIA RECOGNITION THROUGH SPONTANEOUS SPEECH

Topic Editors:
Saturnino Luz, University of Edinburgh, United Kingdom
Fasih Haider, University of Edinburgh, United Kingdom
Davida Fromm, Carnegie Mellon University, United States
Brian MacWhinney, Carnegie Mellon University, United States

Citation: Luz, S., Haider, F., Fromm, D., MacWhinney, B., eds. (2021). Alzheimer's Dementia Recognition Through Spontaneous Speech.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-854-2

Table of Contents

EDITORIAL
published: 21 October 2021
doi: 10.3389/fcomp.2021.780169

Editorial: Alzheimer's Dementia Recognition through Spontaneous Speech

Saturnino Luz[1]*, *Fasih Haider*[1], *Sofia de la Fuente Garcia*[1], *Davida Fromm*[2] and *Brian MacWhinney*[2]

[1] *Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh, United Kingdom,* [2] *Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, United States*

Keywords: Alzheimer's disease, signal processing (SP), machine learning, speech processing and recognition, natural language processing (NLP), computational paralinguistics

OPEN ACCESS

Edited by and reviewed by:
Anton Nijholt,
University of Twente, Netherlands

***Correspondence:**
Saturnino Luz
s.luz@ed.ac.uk

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *20 September 2021*
Accepted: *30 September 2021*
Published: *21 October 2021*

Citation:
Luz S, Haider F, de la Fuente Garcia S, Fromm D and MacWhinney B (2021) Editorial: Alzheimer's Dementia Recognition through Spontaneous Speech. Front. Comput. Sci. 3:780169. doi: 10.3389/fcomp.2021.780169

Editorial on the Research Topic

Alzheimer's Dementia Recognition through Spontaneous Speech

The need for inexpensive, safe, accurate and non-invasive biomarkers for Alzheimer's disease (AD) has motivated much current research (Mandell and Green, 2011). While diagnosis and evaluation of interventions are still primarily done through clinical assessment, "digital biomarkers" have attracted increasing interest. AI-enabled speech and language analysis has emerged as promising such biomarker for the assessment of disease status (de la Fuente Garcia et al., 2020).

While a number of studies have investigated speech and language features for the detection of AD and mild cognitive impairment (Fraser et al., 2016), and proposed various signal processing and machine learning methods for this task (Petti et al., 2020), the field still lacks balanced benchmark data against which different approaches can be systematically compared. This Research Topic addresses this issue by exploring the use of speech characteristics for AD recognition using balanced data and shared tasks, such as those provided by the ADReSS Challenges (Luz et al., 2020, Luz et al., 2021). These tasks have brought together groups working on this active area of research, providing the community with benchmarks for comparison of speech and language approaches to cognitive assessment. Reflecting the multidisciplinary character of the topic, the articles in this collection span three journals: Frontiers of Aging Neuroscience, Frontiers of Computer Science and Frontiers in Psychology.

Most papers in this Reseach Topic target two main tasks: AD classification, for distinguishing individuals with AD from healthy controls, and cognitive test score regression, to infer the patient's Mini Mental Status Examination (MMSE) score (Folstein et al., 1975). Of the twenty papers published in this collection, 14 used the ADReSS dataset (Luz et al., 2020), by itself or in combination with other data. The ADReSS dataset is a curated subset of DementiaBank's Pitt Corpus, matched for age and gender so as to minimise risk of bias in the prediction tasks. The data consist of audio recordings of picture descriptions elicited from participants using the Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Becker et al., 1994; Goodglass et al., 2001), transcribed and annotated using the CHAT coding system (MacWhinney, 2021). The papers covered a variety of approaches and models.

Antonsson et al. aimed to distinguish progressive cognitive decline from stable cognitive impairment using semantic analysis of a discourse task. Support Vector Machine (SVM) models performed best (AUC = 0.93) with both semantic verbal fluency scores and disfluency features from the discourse task. Discourse analysis revealed significantly greater use of unrelated speech in the progressive cognitive decline group compared with the stable group and healthy controls (HC).

Clarke et al. examined the impact of five different speech tasks (picture description, conversation, overlearned narrative recall, procedural recall, novel narrative retelling) on classification of 50 participants: 25 HC, 13 mild AD, 12 MCI. Linguistic features (n = 286) were automatically extracted from each task and used to train SVMs. Classification accuracy varied across tasks (62–78% for HC vs AD + MCI, 59–90% for HC vs AD, 50–78% for HC vs MCI) as did which features were most important to the classification.

Balagopalan et al. used linguistic and acoustic features derived from ADReSS speech and transcripts. They tuned a pretrained BERT model (Devlin et al., 2018) and compared its features to clinically-interpretable language features. The BERT model outperformed other features and achieved accuracy of 83.33% for AD classification. A ridget regressor with 25 pre-engineered features obtained root mean squared error (RMSE) of 4.56 in MMSE prediction.

Chlasta and Wołk used VGGish, a pretrained a Tensorflow model for audio feature extraction and a custom raw waveform based convolutional neural (CNN), DemCNN, to model the acoustic characteristics of AD speech on the ADreSS dataset. DemCNN provided better results than VGGish (Hershey et al., 2017) and achieved an accuracy of 62.5% using only the acoustic information.

De Looze et al. combined structural MRI, neuropsychological testing and conversational features to explore temporal characteristics of speech in a collaborative referencing task. They investigated associations with cognitive function and volumetry in brain areas known to be affected by MCI and AD. A linear mixed-effect model was built for data of 32 individuals to assess the predictive power of conversational speech features to classify clinical groups. They found that slower speech and slower turn-taking may provide useful markers for early detection of cognitive decline.

Guo et al. emphasized the importance of large normative datasets in training accurate and reliable machine learning models for dementia detection. They incorporated a new corpus of Cookie Theft picture descriptions (HC = 839, NC = 115) from the Wisconsin Longitudinal Study (Herd et al., 2014) to train a BERT model and demonstrated improved performance on the detection task compared with results of the model trained on the ADReSS data alone (82.1% vs 79.8, accuracy, and 92.3 vs 88. 3% AUC).

Haulcy and Glass investigated the use of i-vectors and x-vectors (Snyder et al., 2018), which are acoustic features originally devised for speaker identification, and linguistic features to tackle AD detection and MMSE prediction. The i-vectors and x-vectors were pre-trained on existing datasets unrelated to AD as well as in-domain data. Several classification and regression models were tested, yielding 85.4% accuracy in AD detection with SVM and Random Forests, and 4.56 RMSE with a gradient boosting regressor. Linguistic and acoustic features were modelled separately. The former yielded better performance. The authors speculate that the poor performance of i-vectors and x-vectors was due to in- and out-of-domain training data mismatch.

Jonell et al. proposed a multimodal analysis of patient behavior to improve early detection of dementia. Their system captured data from clinical interviews using nine different sensor devices which recorded speech, language, facial gestures, motor signs, gaze, pupil dilation, heart rate variability and thermal emission. This information was gathered from 25 patients with AD and later combined with brain scans, psychological tests, speech therapist assessments and other clinical data. They found that multimodality, in combination with the more established biomarkers, improves clinical discrimination.

Laguarta and Subirana present an approach to the identification of different diseases which combines multiple biomarkers (features), including vocal cords, sentiment, lung and respiratory tract, among others. The authors employed transfer learning from other (non-AD) audio datasets to learn these features. The resulting model achieved up to 93% accuracy on the ADReSS dataset. Interestingly, the respiratory tract features, which were previously used in the detection of COVID-19 from a cough dataset, also proved helpful in AD detection.

Lindsay et al. investigated spontaneous speech of 78 HC and 76 AD individuals in English and French, proposing a multilingual model. Task-specific, semantic, syntactic and paralinguistic features were analysed. They found that language features, excluding task specific features, represent "generalisable" signs for cognitive language impairment in AD, outperforming all other feature sets. Semantic features were the most generalizable, with paralinguistic features showing no overlap between languages.

The work of Mahajan and Baths tested several acoustic and linguistic models, comparing their performance on ADReSS and a larger subset of DementiaBank. They employed a deep learning bimodal model to combine these features. For linguistic models, accuracy was lower on ADReSS than on DementiaBank (73 vs 88%). The authors attribute this to the smaller size of ADReSS and to overfitting in DementiaBank due to repeated samples from the same participant. Although the best linguistic model performed similarly to the bimodal learner, the authors suggest a number of possible improvements.

Martinc et al. presented a multimodal approach to AD detection using ADreSS data. The Active Data Representation method (Haider et al., 2020) was used for fusion of acoustic and textual features at sentence and word level, along with temporal aspects of linguistic features. They achieved an accuracy of 93.75% through late fusion of acoustic, text and temporal models.

Meghanani et al. compared two approaches to the challenge tasks based on use of the non-automatic, hand-created transcripts. Both methods relied on the extraction of n-grams of varying lengths (n = 2,3,4, and 5) from the transcripts. The first method employed CNNs with a single convolutional layer in which the kernel size was adapted to the n-gram size. The second method used the fastText model with bigrams and trigrams. The fastText models outperformed the CNN models, achieving 83.3% accuracy for classification and RMSE of 4.87 for prediction of MMSE scores.

Millington and Luz approached the data representation problem in the ADReSS dataset by converting its text transcriptions into word co-occurrence graphs and computing several graph structure metrics. They found that AD graphs have lower heterogeneity and centralization, but higher edge density. These metrics were used as input features to standard machine learning classifiers and regressors. A graph embedding metric was tested for comparison. Graph metrics outperformed graph embedding, achieving 66.7% accuracy in classification, and a 5.67 RMSE in MMSE regression.

Nasreen et al. investigated the role of conversational features such as dysfluencies, pauses, overlaps and other interactional elements in AD detection. They used the Carolinas Conversations Collection (Pope and Davis, 2011) to create classification models based on those features. The combination of dysfluency and interactional features resulted in a classification accuracy of 90%. These findings in conversational speech seem to agree with the findings from other papers in this Research Topic, which highlighted the importance of pauses and dysfluency in detecting AD in the ADReSS monologue data.

Parvin et al. performed a randomised controlled clinical trial to investigate the effects of dual-task training on 26 patients with AD. Patients performed physical, cognitive and mental assessments and had their brain oscillations measured pre- and post-intervention, which consisted of a 12-weeks visual training program. The trained group showed significant improvements in cognitive function, mood and fitness. This was associated with a significant positive change in brain oscillation.

Sadeghian et al. examined the potential of an almost fully automated system for AD detection. Rather than using DementiaBank, they collected 72 new samples (26 AD, 46 HC) with higher quality audio. ASR was performed on data with pauses removed using voice activity detection. From this, they extracted 236 textual features and then used a genetic algorithm as well as a Multi-Layer Perceptron to identify the 10 most useful features, achieving 94% accuracy in detection.

Shah et al. used speech samples from the DementiaBank database for binary classification and MMSE regression. Although they developed models that combined acoustic and language-based features, their best performing model for binary classification used language-based features only with a regularized logistic regression, achieving 85.4% accuracy on a hold-out test set. A more reduced set of language features was their best performing model for the regression task, with an RMSE of 5.62.

Yuan et al. presented a method for encoding filled and unfilled pauses in transcripts to fine tune the training of language models using BERT and ERNIE. The accuracy of dementia detection improved to 89.6% (with ERNIE). Compared with controls, the individuals with dementia vocalised filled pause *um* much less frequently than *uh*, and their language samples included more pauses.

Zhu et al. used a transfer learning technique to fine-tuning the last layers of a pretrained model with customized layers for AD detection. The MobileNet and YAMNet network architectures were employed for this. They then used speech and text versions of BERT, individually and in combination for the same task. The text models outperformed the speech models, with the version based on pre-training with the longest input frame achieving 89. 58% accuracy. The models which combined audio and speech data generally performed better than the models separately.

The studies in this Research Topic represent the state of the art in dementia detection, and contribute to the increasing body of evidence supporting machine learning and spoken language for detecting cognitive decline.

AUTHOR CONTRIBUTIONS

All authors made substantial contributions to the work and approved this manuscript for publication.

FUNDING

This work funded by the European Union's Horizon 2020 research and innovation programme, grant agreement No 769661 (SAAM project). The original acquisition of the DementiaBank data was supported by NIH grants AG005133 and AG003705 to the University of Pittsburgh.

REFERENCES

Becker, J. T., Boller, F., Lopez, O., Saxton, J., and McGonigle, K. (1994). The Natural History of Alzheimer's Disease. *Arch. Neurol.* 51, 585–594. doi:10.1001/archneur.1994.00540180063015

de la Fuente Garcia, S., Ritchie, C. W., and Luz, S. (2020). Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *J. Alzheimers Dis.* 78, 1547–1574. doi:10.3233/JAD-200888

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. [Preprint] arXiv:1810.04805.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental State". *J. Psychiatr. Res.* 12, 189–198. doi:10.1016/0022-3956(75)90026-6

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2015). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *J. Alzheimers Dis.* 49, 407–422. doi:10.3233/JAD-150520

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *BDAE-3: Boston Diagnostic Aphasia Examination*. Third Edition. Philadelphia, PA: Lippincott Williams & Wilkins.

Haider, F., de la Fuente, S., and Luz, S. (2020). An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE J. Sel. Top. Signal. Process.* 14, 272–281. doi:10.1109/jstsp.2019.2955022

Herd, P., Carr, D., and Roan, C. (2014). Cohort Profile: Wisconsin Longitudinal Study (WLS). *Int. J. Epidemiol.* 43, 34–41. doi:10.1093/ije/dys194

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). "Cnn Architectures for Large-Scale Audio Classification," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, United States, March 5–9, 2017, 131–135. doi:10.1109/ICASSP.2017.7952132

Luz, S., Haider, F., Fuente, S. d. l., Fromm, D., and MacWhinney, B. (2020). Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. *Proc. Interspeech* 2020, 2172–2176. doi:10.21437/Interspeech.2020-2571

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2021). "Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge," in Proceedings of Interspeech 2021, Brno, Czechia, August 30–September 3, 2021, 3780–3784. doi:10.21437/Interspeech.2021-1220

MacWhinney, B. (2021). *Tools for Analyzing Talk Part 1: The CHAT Transcription Format*. Pittsburgh, PA: Carnegie Mellon University. Technical Report. doi:10.21415/3mhn-0z89

Mandell, A., and Green, R. (2011). "Alzheimer's Disease," in *Handbook of Alzheimer's Disease*. Editors A. E. Budson and N. W. Kowall (Malden, MA: John Wiley & Sons), 4–91. chap. 1. doi:10.1002/9781444344110.ch1

Petti, U., Baker, S., and Korhonen, A. (2020). A Systematic Literature Review of Automatic Alzheimer's Disease Detection from Speech and Language. *J. Am. Med. Inform. Assoc.* 27, 1784–1797. doi:10.1093/jamia/ocaa174

Pope, C., and Davis, B. H. (2011). Finding a Balance: The Carolinas Conversation Collection. *Corpus Linguist. Linguist. Theory* 7 (1), 143–161.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). "X-vectors: Robust DNN Embeddings for Speaker Recognition," in Procs IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 5329–5333. doi:10.1109/icassp.2018.8461375

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Dual-Task Training Affect Cognitive and Physical Performances and Brain Oscillation Ratio of Patients With Alzheimer's Disease: A Randomized Controlled Trial

Elnaz Parvin[1], *Fatemeh Mohammadian*[2], *Sadegh Amani-Shalamzari*[1]*, *Mahdi Bayati*[3] and *Behnaz Tazesh*[4]

[1]Department of Exercise Physiology, Faculty of Physical Education and Sports Science, Kharazmi University, Tehran, Iran, [2]Department of Neurology, Roozbeh Hospital, Tehran University of Medical Science, Tehran, Iran, [3]Department of Exercise Physiology, Sports Medicine Research Center, Sport Sciences Research Institute, Tehran, Iran, [4]Sports and Exercise Medicine Specialist, Roozbeh Hospital, Tehran University of Medical Sciences, Tehran, Iran

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Tuba Aktürk,
Istanbul Medipol University, Turkey
Mina Kheirkhah,
University Hospital Jena, Germany
Görsev Yener,
Dokuz Eylul University, Turkey

***Correspondence:**
Sadegh Amani-Shalamzari
amani_sadegh@khu.ac.ir

Received: *11 September 2020*
Accepted: *20 November 2020*
Published: *22 December 2020*

Citation:
Parvin E, Mohammadian F, Amani-Shalamzari S, Bayati M and Tazesh B (2020) Dual-Task Training Affect Cognitive and Physical Performances and Brain Oscillation Ratio of Patients With Alzheimer's Disease: A Randomized Controlled Trial.
Front. Aging Neurosci. 12:605317.
doi: 10.3389/fnagi.2020.605317

This study aimed to investigate the effect of 12 weeks of dual-task training on cognitive status, physical performance, and brain oscillation of patients with Alzheimer's disease (AD). Twenty-six AD patients were randomly assigned to two groups, the training group (TG) and control group (CG). TG executed progressive combined exercises with visual stimulation twice a week for 12 weeks. Training included muscle endurance, balance, flexibility, and aerobic exercises with eyes closed and opened. Brain oscillation on electroencephalography (EEG) and a series of physical, cognitive, and mental tests were taken before and post-intervention. There was a significant improvement after training protocol in cognitive function, particularly in short-term and working memory, attention, and executive function ($p < 0.01$). Besides, there were substantial improvements in depression status (GDS scale), aerobic fitness (6 min walking), flexibility (chair sit and reach) functional ability (chair stand, timed up and go test), strength (knee extensions, preacher biceps curl, handgrip) in TG compared to CG. These signs of progress were associated with a significant increase ($p < 0.05$) in the frequency of brain oscillation and a decrease in the theta/alpha ratio. In addition to physical performance, the regular combined training with visual stimulation improves brain health as indicated by improving cognitive function and reducing the theta/alpha ratio.

Clinical Trial Registration: Iranian Registry of Clinical Trials (IRCT) https://www.irct.ir/, identifier IRCT20190504043468N1—August 5, 2020.

Keywords: aging, alpha wave, cognitive performance, dementia, electroencephalography, exercise, physical activity, theta wave

Abbreviations: 1-RM, one-repetition maximum; 6MWT, six-minute walk test; AD, Alzheimer's disease; ANOVA, analysis of variance; BF, body fat; BMI, body mass index; CG, control group; CSR, chair sit and reach; CV, coefficient of variation; EEG, electroencephalography; ES, effect size; GDS, geriatric depression scale; HR, heart rate; IC, internal consistency; ICC, intra-class correlation coefficient; MCI, mild cognitive impairments; MoCA, montreal cognitive assessment; SD, standard deviation; TG, training group; TUG, timed up and go.

INTRODUCTION

Alzheimer's disease (AD) is a chronic neurodegenerative disease, without any known treatment. This disease progressively destroys brain structures, such as the hippocampus and entorhinal cortex, due to the accumulation of pathological forms of amyloid plaques and neurofibrillary tangles (Lane et al., 2018). Consequently, mental functions, including memory and cognition, are lost, leading to a decline in activities of daily living (McGough et al., 2017). In this regard, Burns et al. (2010) reported that reduced lean body mass in AD is associated with brain atrophy and declined brain function, including cognitive performance. In this regard, a positive correlation has been reported between Montreal Cognitive Assessment (MoCA) test and fitness parameters especially muscle strength (Tolea and Galvin, 2016; Xiao et al., 2020).

The electrical activity of the cerebral cortex (brain oscillation) can be recorded *via* electroencephalography (EEG) by placing electrodes on the scalp. The frequency of resting brain oscillation change in AD patients, compared to older individuals or those with mild cognitive impairments (MCIs), considering a decrease in alpha and beta power (Hsiao et al., 2013; Koelewijn et al., 2017) and an increase in theta power (Moretti et al., 2004; Hsiao et al., 2013). These changes are associated with altered cerebral blood flow, cognitive function (Lizio et al., 2011), and occipital gray matter density (Babiloni et al., 2015). Researchers demonstrate that alpha activity is strongly associated with working memory and probably with long-term memory (Başar, 2012; Başar and Güntekin, 2012). It seems that the brain oscillations ratio is important in relation to brain health. The theta/alpha ratio (Fahimi et al., 2017), which is a marker of AD and cognitive impairments, increases in patients with AD, compared to healthy individuals. A study reported a negative correlation ($r = -0.52$) between the theta/alpha ratio and the MoCA test in patients with type 2 diabetes (Bian et al., 2014). In patients with MCI, occipital alpha slowing may lead to AD (Babiloni et al., 2015). Also, the degree of reduction in alpha and beta peak frequencies is correlated with the stage of AD (Moretti et al., 2004; Koelewijn et al., 2017).

Epidemiological evidence suggests exercise training as a non-pharmacological approach to protect against AD (Rao et al., 2014; Huang et al., 2016; Jia et al., 2019; De la Rosa et al., 2020), increase the hippocampus size (Erickson et al., 2011), and increase brain neurogenesis (Liu and Nusslock, 2018). These structural changes are associated with functional improvements, such as improved independence and cognition of AD patients (Jia et al., 2019). Moreover, these exercise-induced brain changes are associated with alterations in the power of brain oscillation. However, to the best of our knowledge, no studies are investigating the effects of physical training on the frequency of brain oscillation in AD. In this regard, Jiang et al. (2019) reported that a 10-week limb exercise training leads to a significant increase in the alpha and beta wave power values in all brain areas of MCI patients which is associated with psychomotor speed and decline in cognitive function. Also, Gutmann et al. (2015) reported that the individual alpha peak frequency remained unchanged after 4 weeks of moderate exercise training in healthy individuals. Also, researchers have reported that acute bouts of exercise increase the power of beta oscillation in the frontal and central areas of the brain, which may indicate an increase in cortical activation (Moraes et al., 2007; Hubner et al., 2018); however, the long-term effects of physical training are unclear.

Researchers have shown that brain activation during exercise (a dual-task exercise) is beneficial for cognitive function (Brustio et al., 2018; Techayusukcharoen et al., 2019). Generally, training with eyes closed and remembering to do specific exercises with several stations are simple mental activities. In this regard, Hutt and Redding (2014) showed that an eyes-closed dance training increased the dynamic balance of ballet dancers, as closing the eyes led to a shift from visual to proprioceptive dependence for balance control. Moreover, researchers have found that closing the eyes activates different areas of the brain, especially the amygdala, which is involved in memory and learning (Marx et al., 2004; Lerner et al., 2009).

According to some researchers, unlike other oscillations, the power of alpha oscillations increases in a resting state with the eyes closed (Kan et al., 2017), whereas it differs when the person focuses on performing activities with the eyes closed. Dual-task exercises can be used to maintain the brain structure and function and improve physical independence in AD patients. Accordingly, eyes-closed exercises can activate the brain areas involved in memory to focus on activities; they may also increase alpha and beta oscillations (Barry et al., 2007).

Overall, AD causes impairments in different physical and mental functions. To the best of our knowledge, this is the first study to assess the effects of combined physical training with visual stimulation on the physical and cognitive functions of patients with AD. It is known that the power of brain oscillation reflects brain changes and that AD increases the theta/alpha ratio. Accordingly, we hypothesized that physical training combined with mental challenge could modify the power of brain oscillations. In this study, we aimed to investigate the effects of combined training with visual stimulation on the theta/alpha ratio, as well as the cognitive and physical health of patients with AD. Also, we aimed to explore the correlations between cognitive performance and fitness performance, as well as brain oscillations.

MATERIALS AND METHODS

Study Design

This randomized clinical trial, with control and parallel groups, phase 2, and single-blind design was conducted on patients with AD. We aimed to investigate the effects of a 12-week dual-task training (low-intensity exercise with eyes open and closed), on the brain oscillation (alpha, beta, and theta), cognitive and physical performances of patients with AD. One week before the study, the participants and their caregivers attended three familiarization sessions, where they were informed about the benefits and potential risks of the study, signed a consent form, and participated in pretests. The block randomization method was applied before the study, and the participants were assigned to two groups, including the training group (TG) and the control group (CG). Brain oscillation, psychological and cognitive status, and physical fitness parameters, including body composition, aerobic capacity, muscle strength, flexibility, and functional abilities, were assessed in familiarization sessions.

Participants

Patients with AD, who were eligible to participate in this study, were recruited from the memory clinic of Roozbeh Hospital in Tehran, Iran. AD patients, with mild dementia and the ability to walk and move independently, were included in this study. A neurologist confirmed the diagnosis of dementia, based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria. Brain imaging and laboratory tests were performed to exclude other causes of dementia. The AD severity was determined, based on the Functional Assessment Staging Test (FAST).

The patients' medications, including choline esterase inhibitors (rivastigmine and donepezil), memantine, and selective serotonin reuptake inhibitors (SSRIs including sertraline, citalopram, and trazodone), were reviewed before recruiting the patients in the study. The medications were not changed during the intervention in terms of type or dosage. Before entering the study, all patients received cardiac consultation to rule out possible cardiac diseases or ischemia. Patients with serious cardiac diseases (e.g., unstable angina and recent myocardial infarct) were excluded.

Thirty-two eligible subjects volunteered to participate in the study, but the data of 26 patients (age: 67.4 ± 8.8 years; height: 165.8 ± 7.8 cm, body mass: 72.7 ± 11.3 kg, BMI 26.5 ± 4.3 kg/m^2), who completed the pre- and post-tests, were finally analyzed. The participants were randomly assigned into two groups, including the TG and the CG. A CONSORT flow diagram of the present study is shown in **Figure 1**. On the other hand, the exclusion criteria were as follows: (1) deterioration of health condition; (2) inability to perform training; (3) lack of interest in continuing training; (4) not completing the posttest; and (5) the physician's decision to exclude the participant from the study. To estimate the number of participants in each group, a sample size calculation was performed using G*Power Software version 3.1.9.6 (Faul et al., 2007) for repeated measure ANOVA, using a rejection criterion of 0.05 and 0.8 (1-beta) power, and large effect (f = 0.5), a minimum of 13 participants need to each group. All research procedures were approved by the Ethic committees for Sport Sciences Research Institute of Iran (approval number: IR.SSRI.REC.1398.037) and were conducted following the Declaration of Helsinki and reported according to CONSORT guidelines (Schulz et al., 2010). The study has been registered in the Iranian Registry of Clinical Trials (IRCT; one of the Primary Registries in WHO Registry Network) with registration number: IRCT20190504043468N1.

Measures

Before and after training, the participants underwent a series of tests. All training sessions and tests were performed at Roozbeh Hospital Medical Center under the supervision of a sports medicine physician.

Cognitive Status

The Montreal Cognitive Assessment (MoCA) test, developed by Nasreddine et al. (2005) for MCI and dementia, evaluates different domains of cognitive functioning. The reliability of this test was 92%, based on Cronbach's alpha, and its internal consistency (IC) was 83% (Sikaroodi et al., 2012). The maximum score of the test is 30, with a score of 26 or higher considered to be normal. This test, which is executed within 10 min, includes different domains: short-term memory (five points); executive function, including Trail Making Test B, Clock Drawing Test, and visuospatial function test (cube copying; five points); attention and working memory (six points); language, including naming, repetition, and fluency (six points); abstraction (similarity; two points); and orientation to time and place (six points). Patients with scores of 26 or higher did not have any cognitive impairments (normal MoCA), whereas patients with scores lower than 26 probably had cognitive impairments.

Depression Questionnaire

The Geriatric Depression Scale (GDS) was used to assess depression in the participants. In this questionnaire, all questions are of similar weight and have a yes/no response format. The maximum score of GDS is 15, and the minimum score is zero, with higher scores indicating more severe depression. This scale is one of the best tools for measuring depression in the elderly and patients with dementia. The sensitivity of 92% and specificity of 89% have been reported for this questionnaire (Bakhtiyari et al., 2014). The validity and reliability of 15-item GDS were measured by Malakouti et al. (2006) in Iran, and the best cut-off point was eight, with 90% sensitivity and 84% specificity (Sikaroodi et al., 2012).

The 15-item GDS captures depressive symptoms over the past week, using a yes/no response format. For 10 items, a positive response ("yes") is given a score of one, and for five items, a negative response ("no") is given a score of zero. Also, five items are reverse-scored (one for "no" and zero for "yes"). The total score of the items ranges from 0 to 15, with higher scores indicating more depressive symptoms. The GDS-15 score has been used as both continuous and categorical variables elsewhere. We used a cut-off score of ≥5 to indicate the presence of clinical depression symptoms (0, GDS-15 score <5 and 1, GDS-15 score ≥5). We also considered the continuous score of GDS-15 as the outcome (Cron et al., 2016; Honjo et al., 2018; Koohsari et al., 2019).

Anthropometric Indices

Body composition indices, including height (stadiometer, Seca 213, Germany), body mass (digital weighing scales, Seca 769, Germany), body mass index (BMI; kg/m^2), and body fat percentage (BF %; InBody S10, Biospace Company Limited, Seoul, South Korea), were assessed in this study.

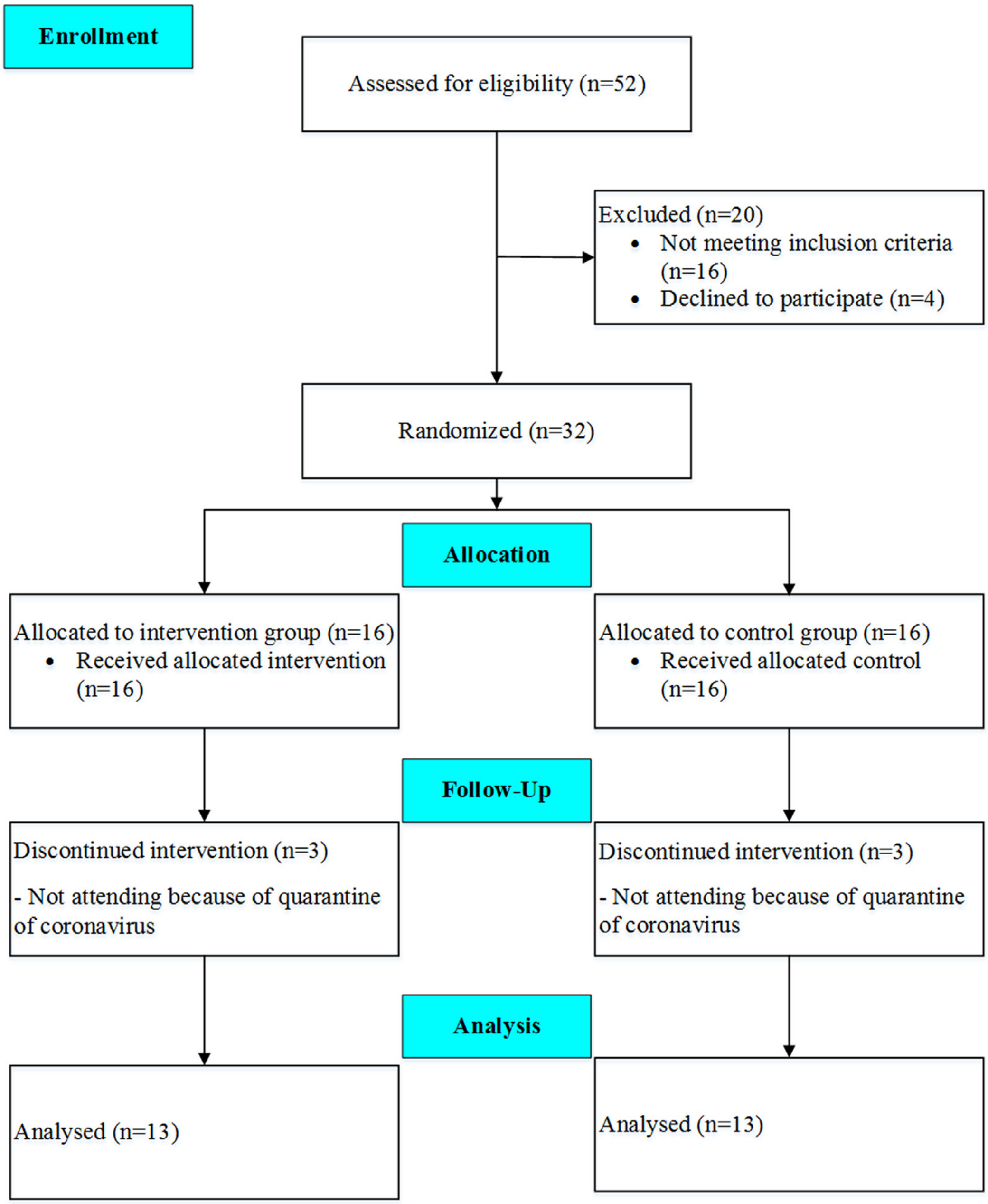

Enrollment
Assessed for eligibility (n=52)
Excluded (n=20)
• Not meeting inclusion criteria (n=16)
• Declined to participate (n=4)
Randomized (n=32)
Allocation
Allocated to intervention group (n=16)
• Received allocated intervention (n=16)
Allocated to control group (n=16)
• Received allocated control (n=16)
Follow-Up
Discontinued intervention (n=3)
- Not attending because of quarantine of coronavirus
Discontinued intervention (n=3)
- Not attending because of quarantine of coronavirus
Analysis
Analysed (n=13)
Analysed (n=13)

FIGURE 1 | CONSORT flow diagram of the study.

Maximum Strength

The maximum strengths of knee extensions, preacher curls, and handgrips were measured for all participants. One-repetition maximum (1-RM) for leg extensions and preacher curls was also determined, based on the procedures described by Sheppard and Triplett (Sheppard and Triplett, 2016). The participants performed a general warm-up, consisting of 5-min pedaling on a stationary bicycle (50–70 rotations per minute at a resistance level of 1–5), followed by a specific warm-up of two sets (5–20 repetitions at 40–50% of perceived maximal effort). Next, they made 3–5 attempts to reach 1RM, with 3–5 min of rest between attempts.

For knee extensions, the participants were asked to sit on a machine (Impulse IT95 Leg Extension, Impulse Health Tech Company Limited, Shandong, China). The researcher adjusted the chair in a way that the subject's legs were placed under the pad, and his/her feet pointed to the pad while extending the knees. In preacher biceps curls, the participant adjusted the preacher bench, held a dumbbell with fully extended arms, and curled it up to shoulder level. Also, a grip strength dynamometer was used to measure the maximum isometric strength of the hand and forearm muscles. After adjusting the handle of the dynamometer for the subjects, they were asked to hold it in their hands, while keeping the arms at the right angles and the elbows on two sides of the body. Participants pressed the dynamometer with maximum isometric effort, which was maintained for about 5 s (Roberts et al., 2011). The best result of the three trials was recorded for each participant.

Functional Tests

The timed up and go (TUG) and chair stand tests were used to measure functional abilities. The TUG test requires the participant to stand up from a chair without the use of arms, walk 2.4 m, turn, return to the chair, and sit (Bigdeli et al., 2020). Also, the chair stand test requires the participant seated on a chair to stand up as many times as possible within 30 s. The participants were instructed to keep their arms crossed at the wrists and hold them in front of the chest. The examiner counted the number of stands performed correctly within 30 s (Rikli and Jones, 2013). Chair "sit-and-reach": (CSR) test requires the participant to sit on the edge of a chair, with one foot flat on the floor and the other leg extended forward with the knee straight and heel on the floor. By placing one hand on top of the other, the subject stretched his/her hands toward the toes by bending at the hip. Next, the distance between the tip of the fingertips and the toes was recorded as a score. If the fingertips reached the toes, the score would be zero; if the fingertips did not touch the toes, the score would be negative; and if the fingertips overlapped, the score would be positive. Overall, two trials were conducted for each participant, and the best distance was recorded (Bigdeli et al., 2020). The six-minute walk test (6MWT) was designed to assess aerobic fitness. In this test, the participants walked at a self-selected pace and were allowed to stop or change their pace (Rikli and Jones, 2013). In the indoor setting, two cones were placed 30 m apart, and the participants were asked to walk back and forth. The walking path was marked every 1 m to determine the distance accurately. For safety, a supervisor accompanied the participants.

Brain Oscillation

Electroencephalography (EEG; SOMNO medics, SSP full EEG, Germany) was used to evaluate the brain oscillation with high sensitivity. The information related to beta, theta, and alpha changes on the EEG test, investigated by a neurologist, was used to determine the patient's status. EEG was obtained over 10 min, and then, the percentage of each brain oscillation and the brain oscillation index were extracted, based on the visual scale. We also divided theta power by alpha power to calculate the theta/alpha ratio.

The 10-20 System which was recommended by the International Federation of Clinical Neurophysiology (IFCN; Deuschl and Eisen, 1999), were used in our study. Also, 21 channels of simultaneous recording are used to obtain EEG recording. In every case, an isolated ground electrode was placed between Cz and Pz.

Interelectrode impedances be checked as a routine prerecording procedure. In our study, impedances up to 10 kOhms are acceptable. Ten seconds of a square wave calibration were made before initiation of the recording in every patient. After that, a visual review of a 30-s run on the system reference montage without the notch filter. The sensitivity of our EEG system was set in 7 μV/mm of trace deflection. The low-frequency filter set in 1 Hz and the high-frequency filter set in 70 Hz to prevent artifacts or changes in electrode impedances that will negatively impact the quality of the EEG.

We record EEG recording at rest in 20 min and then choose 10 min of our recording which has a lower percentage of the artifacts (our patients due to background disease, dementia, had limited and poor cooperation compare to other patients and we should address this point in recording and analyze EEG recording). We reviewed the EEG in at least three different montages including two bipolar and one referential montage. Our recordings included periods when the eyes are open and when they are closed to review the effect of eye-opening on the attenuation of the alpha rhythm. A single-channel electrocardiogram (ECG) is included on one EEG channel.

All EEG recordings were performed in an awake state. According to significant cardiovascular risk factors in numerous dementia patients and patient inability to cooperate, Hyperventilation and Photic stimulation were not performed in patients with AD.

Visual EEG Assessment

The certified clinical neurophysiologists, assessed the entire 20-min EEG recording by visual rating scale and according to a standardized visual rating scheme, which includes the severity of EEG abnormalities and the presence of focal, diffuse, and epileptiform abnormalities.

Source derivation was used as a reference (Hjorth, 1975), and the data was band-pass filtered in four frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz). Oscillations >30 Hz were excluded from further analyses because of the expected artifacts from muscle and eye movement (Hagemann and Naumann, 2001).

Training Protocol

The participants in the experimental group performed 24 workouts twice a week for 12 weeks. Each session lasted about 40–60 min, including 10 min of warm-up, 20–40 min of main exercises, and 10 min of cool down. The participants adhered to a combined protocol, including simple brain activities (eyes-closed training and cognitive activities) and physical activities (muscle endurance, balance, and aerobic capacity). The main training protocol consisted of five parts.

The first part of the training protocol included sitting and standing on an armchair, accompanied by shoulder girdle strengthening (three sets with 5–15 reps, followed by a gradual increase in resistance and repetition, using dumbbells and TheraBand). The second part included crossing over five sponge obstacles (height: 15–20 cm) with eyes closed (two repetitions in the first three sessions, gradually increasing to two reps every three sessions); the distance between the obstacles was variable. In the third part, the participants crossed over a safe balance beam board (2 m) with eyes closed (two repetitions in the first three sessions, gradually increasing to two reps every three sessions).

In the fourth part, six-vowel stations were placed in a semicircular arrangement at a 4-m distance in front of the subject with eyes closed. The subjects were asked to identify the sound of each station, move toward it, perform the predetermined exercises for 15 s (e.g., butterfly curls, Hercules curls, knee raises, hand raises, and biceps curls), and return. There were only two stations in the first session, which increased by one station every three sessions to reach a total of six stations. In the last part, there were four colored lights in front of the participants, each indicating a predetermined exercise. As long as the light was on (10–15 s), the subject was required to perform the relevant exercise (e.g., red light: side-right lunge; blue light: side-left lunge; green light: backward right lunge; and yellow light: backward left lunge). This part lasted for 2 min in the first session, which increased by 1 min every three sessions to reach 5 min by the end.

The exercises changed every three sessions and became more intense. The workouts were performed individually, and each individual attended the center at a certain time. The researcher accompanied the participants throughout the training. The intensity of training was difficult due to the variety of exercises. To monitor the workout intensity, heart rate (HR) was monitored by a smartwatch.

Statistical Methods

Data presented in mean ± standard deviation (SD). The Statistical Package of Social Sciences (SPSS, IBM, v19) was used to analyze data. A repeated measure analysis of variance ANOVA with the time (T1 vs. T2) and protocol (TG vs. CG) was performed to analyze data. To assess the magnitude and direction of the linear correlations between the percentage change of the performance parameters and perceptual indices (MoCA and GDS), bivariate Pearson's correlation coefficient (r) was calculated. Effect size (ES) was also computed as the change score divided by the SD of the change score to examine the magnitude of differences while controlling for the influence of the sample size (Dankel and Loenneke, 2018) with 0.2 considered as a small ES, 0.5 as a moderate ES and >0.8 as a large ES (Batterham and Hopkins, 2006). The significance level was set at $p \leq 0.05$ for all statistical analyses. To determine the test-retest absolute and relative reliability, the coefficient of variation (CV) and intra-class correlation coefficient (ICC) was calculated. The ICC was calculated by a two-way single measure absolute agreement model and the CV was calculated by the formula (CV = [SD/mean] × 100). The CV for tests was <4.0% and ICC was >0.98. Figures were prepared in GraphPad Prism (Version 7.03, GraphPad Software).

RESULTS

Cognitive Performance

The statistical analysis indicated there was a significant main group (between group; $F_{(1,12)}$ = 13.5 p = 0.003, η_p^2: 0.53), time (within group; $F_{(1,12)}$ = 28.1 p = 0.001, η_p^2: 0.70), and interaction effect (group × time; $F_{(1,12)}$ = 40.5 p = 0.001, η_p^2: 0.77) for MoCA. In details, we observed a significant main group ($F_{(1,12)}$ = 7.9 p = 0.016, η_p^2: 0.40), time ($F_{(1,12)}$ = 5.0 p = 0.044, η_p^2: 0.30), and interaction effect ($F_{(1,12)}$ = 13.6 p = 0.003, η_p^2: 0.53) for attention and working memory. We found no significant main group effect for the short-term memory ($F_{(1,12)}$ = 3.2 p = 0.101, η_p^2: 0.21), through a significant time ($F_{(1,12)}$ = 12.9 p = 0.004, η_p^2: 0.52) and interaction effect observed ($F_{(1,12)}$ = 27.0 p = 0.001, η_p^2: 0.69). Also, we observed no significant main group ($F_{(1,12)}$ = 0.1 p = 0.991, η_p^2: 0.01) for the executive function and visuospatial power, through a significant time ($F_{(1,12)}$ = 22.8 p = 0.001, η_p^2: 0.66) and interaction effect existed ($F_{(1,12)}$ = 38.8 p = 0.001, η_p^2: 0.76). However, for orientation, there were no significant main group ($F_{(1,12)}$ = 0.6 p = 0.468, η_p^2: 0.05), time ($F_{(1,12)}$ = 0.7 p = 0.436, η_p^2: 0.05), and interaction effect ($F_{(1,12)}$ = 2.2 p = 0.165, η_p^2: 0.15). In addition, we found no significant main group ($F_{(1,12)}$ = 4.5 p = 0.055, η_p^2: 0.27), time ($F_{(1,12)}$ = 0.23 p = 0.636, η_p^2: 0.02), and interaction effect ($F_{(1,12)}$ = 3.8 p = 0.075, η_p^2: 0.24) for language. Furthermore, there was no significant main group ($F_{(1,12)}$ = 0.02 p = 0.901, η_p^2: 0.01), time ($F_{(1,12)}$ = 3.3 p = 0.096, η_p^2: 0.21), and interaction effect ($F_{(1,12)}$ = 1.9 p = 0.190, η_p^2: 0.14) for the abstraction (**Figure 2**).

Psychological Status

There was no significant main group ($F_{(1,12)}$ = 0.2 p = 0.631, η_p^2: 0.02), but a significant main time ($F_{(1,12)}$ = 23.7 p = 0.001, η_p^2: 0.66) and interaction effect existed ($F_{(1,12)}$ = 21.2, p = 0.001, η_p^2: 0.64) for GDS.

Physical Performance

Descriptive statistics of performance and perceptual parameters pre- and post-intervention are summarized in **Table 1**. In overall, TG compare to CG demonstrated substantial improvements in all performance indices following a 12-week intervention. We found a significant main group ($F_{(1,12)}$ = 6.4 p = 0.026, η_p^2: 0.35), time ($F_{(1,12)}$ = 40.0 p = 0.001, η_p^2: 0.77), and interaction effect ($F_{(1,12)}$ = 53.7 p = 0.001, η_p^2: 0.82) for 6 min walking. For chair sit and reach, there was no significant main group ($F_{(1,12)}$ = 0.9 p = 0.342, η_p^2: 0.07), though a significant time ($F_{(1,12)}$ = 87.6 p = 0.001, η_p^2: 0.88) and interaction effect existed ($F_{(1,12)}$ = 135.9 p = 0.001, η_p^2: 0.92). Furthermore, following the 12-week intervention, we found a significant main group ($F_{(1,12)}$ = 11.2 p = 0.006, η_p^2: 0.48), time ($F_{(1,12)}$ = 80.2 p = 0.001, η_p^2: 0.87), and interaction effect ($F_{(1,12)}$ = 61.3 p = 0.001, η_p^2: 0.84) for strength of preacher biceps curl. For strength of knee extensions, there also was a significant main group ($F_{(1,12)}$ = 6.1 p = 0.030, η_p^2: 0.34), time ($F_{(1,12)}$ = 25.1 p = 0.001, η_p^2: 0.68), and interaction effect ($F_{(1,12)=}$ 38.2 p = 0.001, η_p^2: 0.76). For strength of handgrip, there was no significant main group ($F_{(1,12)}$ = 2.3 p = 0.152, η_p^2: 0.16), but significant time ($F_{(1,12)}$ = 63.6 p = 0.001, η_p^2: 0.84) and interaction effect observed ($F_{(1,12)}$ = 74.2 p = 0.001, η_p^2: 0.86).

For functional indices, we found a significant main group ($F_{(1,12)}$ = 12.7 p = 0.004, η_p^2: 0.52), time ($F_{(1,12)}$ = 90.9 p = 0.001, η_p^2: 0.88), and interaction effect ($F_{(1,12)}$ = 172.1 p = 0.001, η_p^2: 0.94) for timed up and go test. In addition, there was a significant main group ($F_{(1,12)}$ = 29.0 p = 0.001, η_p^2: 0.71), time ($F_{(1,12)}$ = 54.6 p = 0.001, η_p^2: 0.82), and interaction effect ($F_{(1,12)}$ = 41.1 p = 0.001, η_p^2: 0.77) for chair stand.

TABLE 1 | Performance and psychological characteristics of participants pre- and post-intervention.

Variable	Group	Pre	Post	% change	Cohen's *d*	*p*
6 min walking (m)	TG	177.0 ± 81.5	318.9 ± 85.5	96.9	1.9	0.001
	CG	180.0 ± 66.2	174.2 ± 62.9	−2.6	−0.5	
Knee extension (kg)	TG	10.8 ± 5.6	23.1 ± 10.6	134.6	1.6	0.001
	CG	10.7 ± 4.9	10.0 ± 4.5	−4.5	−0.6	
Biceps curl (kg)	TG	6.4 ± 1.7	10.6 ± 2.5	70.2	2.6	0.001
	CG	6.4 ± 1.6	6.1 ± 1.5	−3.1	−0.3	
Handgrip (kg)	TG	23.1 ± 9.1	31.6 ± 8.9	47.9	2.4	0.001
	CG	22.61 ± 8.3	21.3 ± 8.1	−7.1	−1.4	
30 s stand-up (N)	TG	10.3 ± 3.4	18.5 ± 0.37	94.8	1.9	0.001
	CG	9.7 ± 3.1	9.2 ± 2.3	−2.7	−0.5	
Timed Up and Go test (s)	TG	11.8 ± 2.5	6.4 ± 1.4	−45.6	−3.5	0.001
	CG	11.8 ± 2.7	12.4 ± 2.9	5.7	0.8	
Chair sit and reach (cm)	TG	18.5 ± 8.1	26.9 ± 7.6	54.5	3.9	0.001
	CG	19.8 ± 8.1	18.7 ± 7.1	−3.7	−0.5	
MoCA	TG	18.6 ± 3.5	23.9 ± 2.3	28.4	1.7	0.001
	CG	19.0 ± 2.1	17.9 ± 2.2	−3.3	−0.5	
GDS	TG	5.4 ± 2.9	2.6 ± 1.9	−49.5	−1.4	0.001
	CG	4.4 ± 2.5	4.5 ± 2.5	6.8	0.2	
Alpha oscillation (%)	TG	80.0 ± 5.5	83.2 ± 2.7	4.3	0.7	0.002
	CG	78.8 ± 7.7	78.0 ± 7.6	−1.0	−0.6	
Beta oscillation (%)	TG	3.7 ± 3.1	8.9 ± 3.3	218.3	2.6	0.001
	CG	3.1 ± 1.6	3.0 ± 1.1	5.0	−0.2	
Theta oscillation (%)	TG	16.2 ± 6.5	7.8 ± 3.7	−51.8	−1.9	0.001
	CG	18.1 ± 6.9	19.0 ± 7.0	5.9	0.6	

TG, training group; CG, control group; MoCA, the montreal cognitive assessment; GDS, geriatric depression scale.

Brain Oscillation

Following 12 weeks of combined training, the percentage of resting average frequency of brain oscillation in occipital region in the TG increased significantly by 14.5%; change from alpha range to beta frequency (11.51 to 13.15 Hz), but there was no significant change (−1.4%) in control group (11.13 to 10.95 Hz). Descriptive statistics of the brain oscillation are presented in **Table 1**. The results of repeated measure ANOVA showed there was a significant main group ($F_{(1,12)} = 11.4$ $p = 0.005$, η^2_p: 0.48), time ($F_{(1,12)} = 63.7$ $p = 0.001$, η^2_p: 0.84), and interaction effects ($F_{(1,12)} = 39.7$ $p = 0.001$, η^2_p: 0.77) for resting average frequency of brain oscillation (**Figure 3**). We found no significant main group ($F_{(1,12)} = 3.2$ $p = 0.098$, η^2_p: 0.21) and time ($F_{(1,12)} = 3.6$ $p = 0.080$, η^2_p: 0.23) effect, though a significant interaction effect existed ($F_{(1,12)} = 6.7$ $p = 0.024$, η^2_p: 0.36) for percentage of alpha oscillation. While for percentage of beta oscillation, a significant main group ($F_{(1,12)} = 19.2$ $p = 0.001$, η^2_p: 0.62), time ($F_{(1,12)} = 77.2$ $p = 0.001$, η^2_p: 0.86), and interaction effect ($F_{(1,12)} = 82.1$ $p = 0.001$, η^2_p: 0.87) was observed. For percentage of theta oscillation, there was a significant main group ($F_{(1,12)} = 14.7$ $p = 0.002$, η^2_p: 0.55), time ($F_{(1,12)} = 39.5$ $p = 0.001$, η^2_p: 0.77), and interaction effect ($F_{(1,12)} = 46.2$ $p = 0.001$, η^2_p: 0.79). There was a significant main group ($F_{(1,12)} = 10.5$ $p = 0.007$, η^2_p: 0.47), time ($F_{(1,12)} = 29.1$ $p = 0.001$, η^2_p: 0.71) and interaction effect ($F_{(1,12)} = 33.7$ $p = 0.001$, η^2_p: 0.74) for theta/alpha ratio (**Figure 3**).

Correlations

Table 2 presents the bivariate Pearson's correlation coefficient (*r*) between the percentage change of performance parameters and MoCA and GDS. In general, there were moderate to large, positive correlations between MoCA changes and performance induces. Moderate, negative correlations were found between changes in GDS and performance indices. Also, MoCA correlated negatively with the theta/alpha ratio, while GDS correlated positively.

Exercise Monitoring

The mean (SD) of HR during the intervention period was presented in **Figure 4**. The training began at 50% of maximal HR and reached 70% of maximal HR toward the end of the intervention. The range of HR was 80–125 beat per minute.

DISCUSSION

This study aimed to evaluate the efficacy of a 12-week of combined training intervention with visual stimulation on the frequency of brain oscillation, cognitive status, and physical performance of patients with AD. The results revealed that following the intervention, patients in the TG group experienced significant improvements in cognitive function, particularly short-term and working memory, attention, and executive function. We also found significant improvements in the depression status of the TG group, compared to CG.

Moreover, significant improvements were observed in the overall physical performance of the participants. These improvements were paralleled with the reduction of the theta/alpha ratio, suggesting that the intervention was effective in involving and activating neurons. Also, moderate to relatively strong correlations were observed between cognitive and performance indices. The findings of our study revealed that the combination of exercise training with mental challenges (such as closing the eyes, attending to auditory stimuli, and trying to

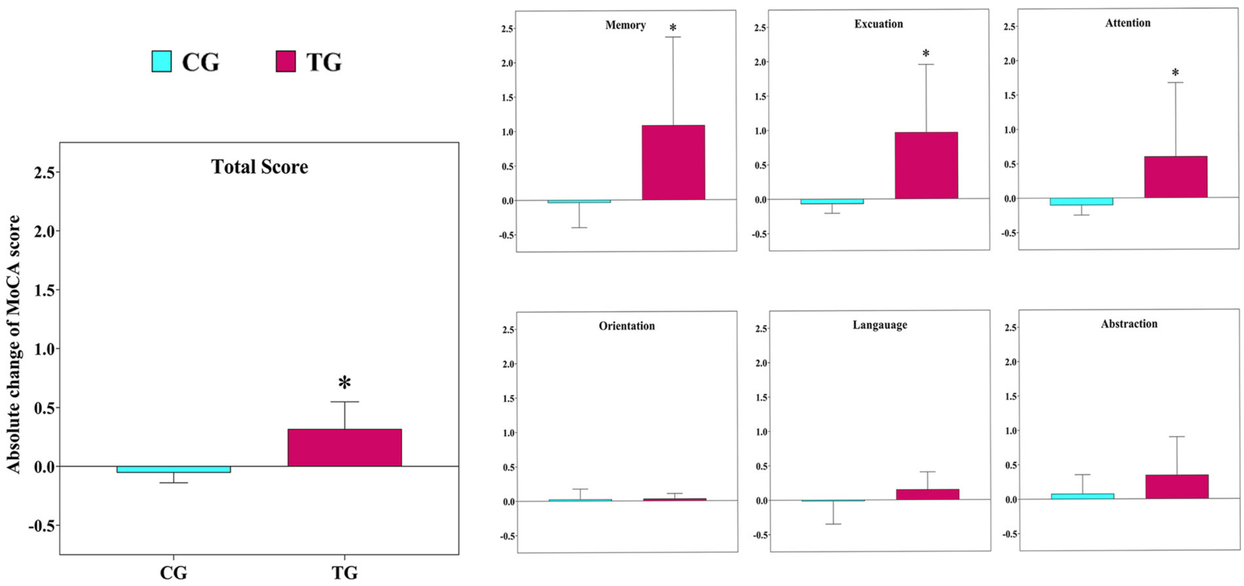


FIGURE 2 | The absolute changes in the scores of the Montreal Cognitive Assessment (MoCA) test following the 12-week intervention in both groups. TG, training group; CG, control group. *The significant difference between groups.

control balance by relying on proprioceptive receptors) can be used to improve the independence of patients with AD.

Cognitive impairments, including memory, speech, attention, and executive function impairments, are among the characteristics of AD, which can be measured with the MoCA test in this population. In our study, after the intervention, cognitive performance (the MoCA test) improved with a large effect size (28.4%; η_p^2 = 1.7). Improvements were observed in short-term memory, executive function, attention, and working memory. Also, since closing eyes activate different areas of the brain, in particular the hippocampus (Ben-Simon et al., 2008), which plays roles in spatial memory, balance, and concentration (Rubin et al., 2014), and since our subjects had to close their eyes during the training protocol, our results might be related to hippocampal activation. However, we have not demonstrated this in our study and suggest it for further investigation in this area. Our results are in agreement with previous research, supporting the protective effects of physical training on cognitive function (Burns et al., 2008; Morris et al., 2017).

Although the exact mechanisms of the protective effect of exercise training on the mental health of AD patients are less clear, several mechanisms have been proposed, including the increase of blood supply to the brain, improvement of metabolic health, production of neurotrophic factors (Gallaway et al., 2017), increased size of the hippocampus (Erickson et al., 2011), and increasing gray and white matter volumes in the inferior parietal cortex and the hippocampus over a long period (Burns et al., 2008; Voss et al., 2013). These alterations are associated with memory and cognitive performance, as well as changes in the power of brain oscillation. In this regard, a previous study showed that even a 12-week period of aerobic training could expedite neuroplasticity and promote brain health in sedentary adults (Chapman et al., 2013); the observed improvements in brain function were attributed to the increased physical activity of the participants.

Depression is one of the most common symptoms and consequences of AD, which exacerbates the negative consequences of this disease. Research has shown that regular exercise training in the short-term had obvious effects on depression management (Craft and Perna, 2004). The results of the present study also demonstrated the effectiveness of combined training in reducing depression. Based on the results, depression was inversely correlated with physical fitness indices and positively correlated with the theta/alpha ratio. Several mechanisms can justify the positive effects of exercise training on depression. Improvement of independence, daily life activities, and the mood is among the advantages of exercise training for reducing depression. Also, social interaction between participants in the TG during the training period was effective in improving mood and managing depression.

Moreover, exercise-mediated production of neurotransmitters, such as dopamine, serotonin (Paillard et al., 2015), and brain-derived neurotrophic factor (Wang and Holsinger, 2018), contributes to the treatment of depression. Also, AD-induced high cortisol levels exert neurotoxic effects on the hippocampus and promote oxidative stress, leading to depression, neurodegeneration, and cognitive decline (Ouanes and Popp, 2019). On the other hand, one of the protective effects of regular exercise is lowering the serum cortisol level (Corazza et al., 2014). Although these factors were not measured in this study, the observed improvements can be explained by these mechanisms.

Researches showed that the changes in the ratio of alpha, beta, and theta oscillations are the AD markers, so we extracted the data of these brain oscillations. On the other hand, we did not consider the gamma and delta oscillations, because the

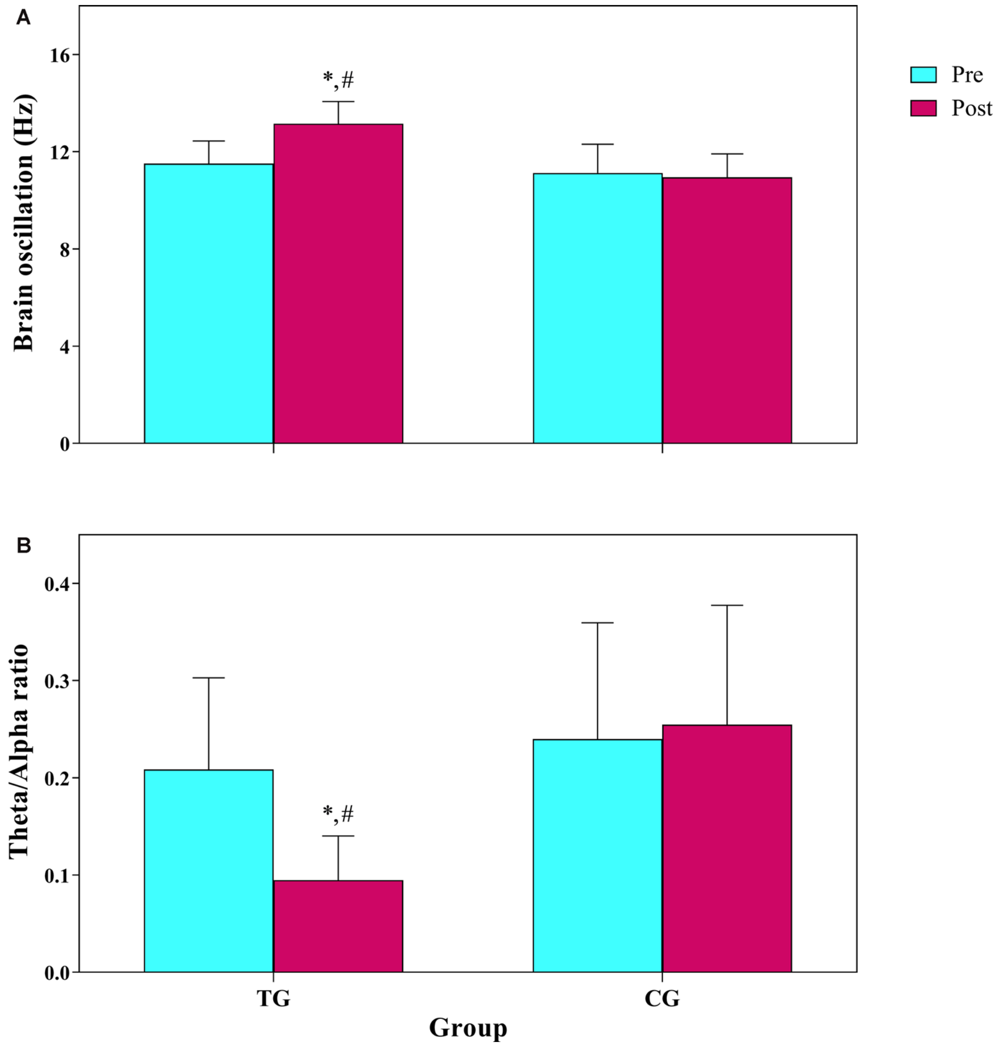


FIGURE 3 | (A) Frequency of brain oscillation, **(B)** theta/alpha ratio in 10 min resting EEG. TG, training group; CG, control group. *The significant difference with pre-test. #The significant difference between groups.

TABLE 2 | Pearson's correlation coefficient between the variables.

		Handgrip	Knee extension	Biceps curl	30 s stand-up	Timed Up and Go test	6 min walking	Chair sit and reach	Theta/alpha ratio
MoCA	*r*	0.81	0.78	0.63	0.68	−0.68	0.36	0.74	−0.56
	p	0.001	0.001	0.001	0.001	0.001	0.067	0.001	0.003
GDS	*r*	−0.50	−0.55	−0.61	−0.203	0.56	−0.57	−0.47	0.62
	p	0.010	0.004	0.001	0.319	0.003	0.003	0.016	0.001

delta and gamma oscillations are activated during sleep and cognitive learning activities, respectively (Abhang et al., 2016). The resting alpha and beta oscillation indicate relaxed and alert wakefulness (Abhang et al., 2016), and the theta/alpha ratio is indicative of cognitive deficits (Fahimi et al., 2017). Decreased alpha oscillation power has been reported in AD (Hsiao et al., 2013; Koelewijn et al., 2017), which is associated with an increase in the theta/alpha ratio. Therefore, the reduction of theta/alpha ratio in our study suggests that a combined training period with mental challenges for AD patients activates the mechanisms in the brain, which improve cognitive processing. This finding is in line with a previous study, which showed that 10 weeks of limb exercise significantly increase the alpha and beta oscillation power in all brain areas of older adults with MCI (Jiang et al., 2019); however, this study did not report the theta/alpha ratio.

Although the exact mechanisms of change in the brain oscillation ratio due to exercise training are unknown in AD, the alpha oscillation power seems to be correlated with

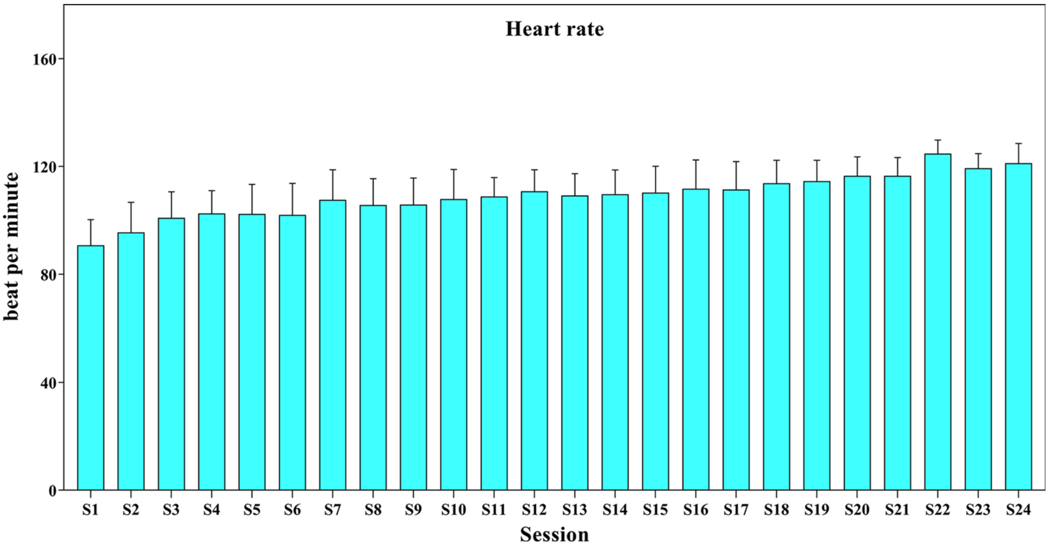


FIGURE 4 | The heart rate (beat/min) during the training sessions.

higher cerebral blood flow in the brain areas, involved in attentional modulation (Jann et al., 2010). Alpha oscillations are generated mainly in the occipital and parietal lobes, as well as thalamocortical feedback loops, whereas beta oscillations mainly originate from the frontal and temporal lobes (Abhang et al., 2016). The eye-closing part of our training protocol forced the individuals to focus on the auditory and proprioceptive data, originating from beta and alpha oscillation.

Moreover, the sensory data are distributed in different areas of the cortex through the thalamus. Therefore, our intervention was highly effective in activating the brain parts involved in attention. In contrast, Gutmann et al. (2015) were reported no changes in alpha oscillation power after 4 weeks of moderate exercise training. It seems that methodological differences can explain these contradicting results. The subjects of the latter study were healthy young men, while the populations of our study were older AD adults. Overall, the findings demonstrated that 12 weeks of training combined with mental challenge reduced the theta/alpha ratio by improving the neurophysiological mechanisms.

AD is associated with the loss of muscle mass and strength, reduced balance, and reduced cardiovascular fitness, leading to inability to perform daily activities, loss of independence, and poor quality of life (Santana-Sosa et al., 2008; Burns et al., 2010; Lane et al., 2018) therefore, our subjects' baseline fitness level was very poor. Our findings showed that 3 months of combined training caused substantial improvements in the performance indices. Resistance exercises (dumbbells, TheraBand, and rubbers) led to increased strength and maintenance of muscle mass, balance exercises (walking on a beam board) and eyes-closed exercises improved proprioception, and consecutive exercises led to increased cardiovascular fitness.

Improved balance in the present study is especially important, as balance and mobility impairments in AD patients are associated with the risk of falling and reduced quality of life. It is worth mentioning that the observed improvements after exercise training are not population-specific, as comparable increments have been observed in the physical capacity of other populations after a short-term training program (de Vreede et al., 2005). Improved fitness components appear to be correlated with the ability to perform daily tasks and quality of life. This finding is in line with a previous study, which examined the effects of exercise training on functional capacity in AD patients (Santana-Sosa et al., 2008).

Santana-Sosa et al. (2008) demonstrated that a 12-week combined training program led to significant improvements in the upper and lower body muscle strength, endurance fitness, balance, and ability to perform daily activities. Also, moderate-to-large positive interactions were observed between changes in physical parameters and cognitive function. Moreover, there was a strong association between the change of muscle strength (especially handgrip) and MoCA. This finding was supported by a previous study, which showed a strong relationship between muscle atrophy and declined cognitive function (Burns et al., 2010). Kim et al. (2019) also reported a positive relationship between the handgrip strength and cognitive function of elderly Korean adults. Moreover, Burns et al. (2008) reported that increased cardiorespiratory fitness is associated with reduced brain atrophy in AD patients. Based on the findings, exercise training can be an important adjunct to the pharmacological treatment of AD.

We acknowledge that there are some limitations to this study. First, the posttest date coincided with the pandemic of COVID-19 in Iran, and we lost some of our participants. Second, due to the lack of full-time caregivers, transportation was difficult, and the workout time was not consistent; however, all participants completed 24 workout sessions. Third, we did not have access to quantitative EEG; therefore, we suggest using structural and functional brain imaging to assess quantitative changes in the brain structure and function in the future. Finally, we did not determine the period when these adaptations remained constant, which indicates the importance of follow-up after 3, 6, or even 12 months.

CONCLUSIONS

In conclusion, a 12-week combined training program, including resistance, balance, and cardiovascular exercises with closed-eyes stimulation, improved the performance capacity of patients with AD. Also, this intervention improved brain health and activated neurophysiological mechanisms, which are associated with increased cognitive function and decreased theta/alpha ratio. Moreover, our findings supported the hypothesis that cognitive functions are correlated with muscle strength-related physical fitness in patients with AD.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethic committees for Sport Sciences Research Institute of Iran with code IR.SSRI.REC.1398.037 and were conducted in accordance with the Declaration of Helsinki. The study has been registered in the Iranian Registry of Clinical Trials (IRCT) with registration number: IRCT20190504043468N1. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SA-S and EP conceived the study. EP, FM, and BT conducted the experiments. SA-S and MB analyzed the study. SA-S, FM, and MB interpreted the data for the study. All authors made substantial contributions to the design of the work, drafted the work or revised it critically for important intellectual content, provided final approval of the version to be published, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All authors read and approved the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ACKNOWLEDGMENTS

We would like to acknowledge the memory clinic and sport setting of Roozbeh Hospital, and as well as the patients and caregivers who collaborated with us. This manuscript has been released as a pre-print at ResearchSquare (Parvin et al., 2020).

REFERENCES

Abhang, P. A., Bharti, W. G., and Suresh, C. M. (2016). *Introduction to EEG-and Speech-Based Emotion Recognition.* London: Academic Press.

Babiloni, C., Del Percio, C., Boccardi, M., Lizio, R., Lopez, S., Carducci, F., et al (2015). Occipital sources of resting-state alpha rhythms are related to local gray matter density in subjects with amnesic mild cognitive impairment and Alzheimer's disease. *Neurobiol. Aging* 36, 556–570. doi: 10.1016/j.neurobiolaging.2014.09.011

Bakhtiyari, F., Foroughan, M., Fakhrzadeh, H., Nazari, N., Najafi, B., Alizadeh, M., et al. (2014). Validation of the persian version of Abbreviated Mental Test (AMT) in elderly residents of Kahrizak charity foundation. *Iran. J. Diabetes Metab.* 13, 487–494.

Barry, R. J., Clarke, A. R., Johnstone, S. J., Magee, C. A., and Rushb, J. A. (2007). EEG differences between eyes-closed and eyes-open resting conditions. *Clin. Neurophysiol.* 118, 2765–2773. doi: 10.1016/j.clinph.2007.07.028

Başar, E. (2012). A review of alpha activity in integrative brain function: fundamental physiology, sensory coding, cognition and pathology. *Int. J. Psychophysiol.* 86, 1–24. doi: 10.1016/j.ijpsycho.2012.07.002

Başar, E., and Güntekin, B. (2012). A short review of alpha activity in cognitive processes and in cognitive impairment. *Int. J. Psychophysiol.* 86, 25–38. doi: 10.1016/j.ijpsycho.2012.07.001

Batterham, A. M., and Hopkins, W. G. (2006). Making meaningful inferences about magnitudes. *Int. J. Sports Physiol. Perform.* 1, 50–57. doi: 10.1123/ijspp.1.1.50

Ben-Simon, E., Podlipsky, I., Arieli, A., Zhdanov, A., and Hendler, T. (2008). Never resting brain: simultaneous representation of two alpha related processes in humans. *PLoS One* 3:e3984. doi: 10.1371/journal.pone.0003984

Bian, Z., Li, Q., Wang, L., Lu, C., Yin, S., and Li, X. (2014). Relative power and coherence of EEG series are related to amnestic mild cognitive impairment in diabetes. *Front. Aging Neurosci.* 6:11. doi: 10.3389/fnagi.2014.00011

Bigdeli, S., Dehghaniyan, M. H., Amani-Shalamzari, S., Rajabi, H., and Gahreman, D. E. (2020). Functional training with blood occlusion influences muscle quality indices in older adults. *Arch. Gerontol. Geriatr.* 90:104110. doi: 10.1016/j.archger.2020.104110

Brustio, P. R., Rabaglietti, E., Formica, S., and Liubicich, M. E. (2018). Dual-task training in older adults: the effect of additional motor tasks on mobility performance. *Arch. Gerontol. Geriatr.* 75, 119–124. doi: 10.1016/j.archger.2017.12.003

Burns, J. M., Cronk, B. B., Anderson, H. S., Donnelly, J. E., Thomas, G. P., Harsha, A., et al. (2008). Cardiorespiratory fitness and brain atrophy in early Alzheimer disease. *Neurology* 71, 210–216. doi: 10.1212/01.wnl.0000317094.86209.cb

Burns, J. M., Johnson, D. K., Watts, A., Swerdlow, R. H., and Brooks, W. M. (2010). Reduced lean mass in early Alzheimer disease and its association with brain atrophy. *Arch. Neurol.* 67, 428–433. doi: 10.1001/archneurol.2010.38

Chapman, S. B., Aslan, S., Spence, J. S., Defina, L. F., Keebler, M. W., Didehbani, N., et al (2013). Shorter term aerobic exercise improves brain, cognition and cardiovascular fitness in aging. *Front. Aging Neurosci.* 5:75. doi: 10.3389/fnagi.2013.00075

Corazza, D. I., Sebastião, É., Pedroso, R. V., Almeida Andreatto, C. A., de Melo Coelho, F. G., Gobbi, S., et al (2014). Influence of chronic exercise on serum cortisol levels in older adults. *Eur. Rev. Aging Phys. Act.* 11, 25–34. doi: 10.1007/s11556-013-0126-8

Craft, L. L., and Perna, F. M. (2004). The benefits of exercise for the clinically depressed. *Prim. Care Companion J. Clin. Psychiatry* 6, 104–111. doi: 10.4088/pcc.v06n0301

Cron, D. C., Friedman, J. F., Winder, G. S., Thelen, A. E., Derck, J. E., Fakhoury, J. W., et al (2016). Depression and frailty in patients with end-stage liver disease referred for transplant evaluation. *Am. J. Transplant.* 16, 1805–1811. doi: 10.1111/ajt.13639

Dankel, S. J., and Loenneke, J. P. (2018). Effect sizes for paired data should use the change score variability rather than the pre-test variability. *J. Strength Cond. Res.*. doi: 10.1519/jsc.0000000000002946. [Epub ahead of print].

De la Rosa, A., Olaso-Gonzalez, G., Arc-Chagnaud, C., Millan, F., Salvador-Pascual, A., García-Lucerga, C., et al. (2020). Physical exercise in the prevention and treatment of Alzheimer's disease. *J. Sport Health Sci.* 9, 394–404. doi: 10.1016/j.jshs.2020.01.004

de Vreede, P. L., Samson, M. M., van Meeteren, N. L., Duursma, S. A., and Verhaar, H. J. (2005). Functional-task exercise versus resistance strength exercise to improve daily function in older women: a randomized, controlled trial. *J. Am. Geriatr. Soc.* 53, 2–10. doi: 10.1111/j.1532-5415.2005.53003.x

Deuschl, G., and Eisen, A. (1999). *Recommendations For the Practice of Clinical Neurophysiology: Guidelines of the International Federation of Clinical Neurophysiology.* 2nd Edn. Vol. Supplement 52. Amsterdam: Elsevier.

Erickson, K. I., Voss, M. W., Prakash, R. S., Basak, C., Szabo, A., Chaddock, L., et al. (2011). Exercise training increases size of hippocampus and improves memory. *Proc. Natl. Acad. Sci. U S A* 108, 3017–3022. doi: 10.1073/pnas.1015950108

Fahimi, G., Tabatabaei, S. M., Fahimi, E., and Rajebi, H. (2017). Index of theta/alpha ratio of the quantitative electroencephalogram in Alzheimer's disease: a case-control study. *Acta Med. Iran* 55, 502–506.

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/bf03193146

Gallaway, P. J., Miyake, H., Buchowski, M. S., Shimada, M., Yoshitake, Y., Kim, A. S., et al. (2017). Physical activity: a viable way to reduce the risks of mild cognitive impairment, Alzheimer's disease and vascular dementia in older adults. *Brain Sci.* 7:22. doi: 10.3390/brainsci7020022

Gutmann, B., Mierau, A., Hulsdunker, T., Hildebrand, C., Przyklenk, A., Hollmann, W., et al. (2015). Effects of physical exercise on individual resting state EEG alpha peak frequency. *Neural Plast.* 2015:717312. doi: 10.1155/2015/717312

Hagemann, D., and Naumann, E. (2001). The effects of ocular artifacts on (lateralized) broadband power in the EEG. *Clin. Neurophysiol.* 112, 215–231. doi: 10.1016/s1388-2457(00)00541-1

Hjorth, B. (1975). An on-line transformation of EEG scalp potentials into orthogonal source derivations. *Electroencephalogr. Clin. Neurophysiol.* 39, 526–530. doi: 10.1016/0013-4694(75)90056-5

Honjo, K., Tani, Y., Saito, M., Sasaki, Y., Kondo, K., Kawachi, I., et al. (2018). Living alone or with others and depressive symptoms and effect modification by residential social cohesion among older adults in Japan: the JAGES longitudinal study. *J. Epidemiol.* 28, 315–322. doi: 10.2188/jea.JE20170065

Hsiao, F.-J., Wang, Y.-J., Yan, S.-H., Chen, W.-T., and Lin, Y.-Y. (2013). Altered oscillation and synchronization of default-mode network activity in mild Alzheimer's disease compared to mild cognitive impairment: an electrophysiological study. *PLoS One* 8:e68792. doi: 10.1371/journal.pone.0068792

Huang, P., Fang, R., Li, B. Y., and Chen, S. D. (2016). Exercise-related changes of networks in aging and mild cognitive impairment brain. *Front. Aging Neurosci.* 8:47. doi: 10.3389/fnagi.2016.00047

Hubner, L., Godde, B., and Voelcker-Rehage, C. (2018). Acute exercise as an intervention to trigger motor performance and EEG beta activity in older adults. *Neural Plast.* 2018:4756785. doi: 10.1155/2018/4756785

Hutt, K., and Redding, E. (2014). The effect of an eyes-closed dance-specific training program on dynamic balance in elite pre-professional ballet dancers: a randomized controlled pilot study. *J. Dance Med. Sci.* 18, 3–11. doi: 10.12678/1089-313X.18.1.3

Jann, K., Koenig, T., Dierks, T., Boesch, C., and Federspiel, A. (2010). Association of individual resting state EEG alpha frequency and cerebral blood flow. *NeuroImage* 51, 365–372. doi: 10.1016/j.neuroimage.2010.02.024

Jia, R. X., Liang, J. H., Xu, Y., and Wang, Y. Q. (2019). Effects of physical activity and exercise on the cognitive function of patients with Alzheimer disease: a meta-analysis. *BMC Geriatr.* 19:181. doi: 10.1186/s12877-019-1175-2

Jiang, H., Chen, S., Wang, L., and Liu, X. (2019). An investigation of limbs exercise as a treatment in improving the psychomotor speed in older adults with mild cognitive impairment. *Brain Sci.* 9:277. doi: 10.3390/brainsci9100277

Kan, D. P. X., Croarkin, P. E., Phang, C. K., and Lee, P. F. (2017). EEG differences between eyes-closed and eyes-open conditions at the resting stage for euthymic participants. *Neurophysiology* 49, 432–440. doi: 10.1016/j.clinph.2007.07.028

Kim, K. H., Park, S. K., Lee, D. R., and Lee, J. (2019). The relationship between handgrip strength and cognitive function in elderly koreans over 8 years: a prospective population-based study using korean longitudinal study of ageing. *Korean J. Fam. Med.* 40, 9–15. doi: 10.4082/kjfm.17.0074

Koelewijn, L., Bompas, A., Tales, A., Brookes, M. J., Muthukumaraswamy, S. D., Bayer, A., et al (2017). Alzheimer's disease disrupts alpha and beta-band resting-state oscillatory network connectivity. *Clin. Neurophysiol.* 128, 2347–2357. doi: 10.1016/j.clinph.2017.04.018

Koohsari, M. J., McCormack, G. R., Nakaya, T., Shibata, A., Ishii, K., Yasunaga, A., et al. (2019). Urban design and Japanese older adults' depressive symptoms. *Cities* 87, 166–173. doi: 10.1016/j.cities.2018.09.020

Lane, C. A., Hardy, J., and Schott, J. M. (2018). Alzheimer's disease. *Eur. J. Neurol.* 25, 59–70. doi: 10.1111/ene.13439

Lerner, Y., Papo, D., Zhdanov, A., Belozersky, L., and Hendler, T. (2009). Eyes wide shut: amygdala mediates eyes-closed effect on emotional experience with music. *PLoS One* 4:e6230. doi: 10.1371/journal.pone.0006230

Liu, P. Z., and Nusslock, R. (2018). Exercise-mediated neurogenesis in the hippocampus *via* BDNF. *Front. Neurosci.* 12:52. doi: 10.3389/fnins.2018.00052

Lizio, R., Vecchio, F., Frisoni, G. B., Ferri, R., Rodriguez, G., and Babiloni, C. (2011). Electroencephalographic rhythms in Alzheimer's disease. *Int. J. Alzheimers Dis.* 2011:927573. doi: 10.4061/2011/927573

Malakouti, S. K., Fatollahi, P., Mirabzadeh, A., Salavati, M., and Zandi, T. (2006). Reliability, validity and factor structure of the GDS-15 in Iranian elderly. *Int. J. Geriatr. Psychiatry* 21, 588–593. doi: 10.1002/gps.1533

Marx, E., Deutschländer, A., Stephan, T., Dieterich, M., Wiesmann, M., and Brandt, T. (2004). Eyes open and eyes closed as rest conditions: impact on brain activation patterns. *NeuroImage* 21, 1818–1824. doi: 10.1016/j.neuroimage.2003.12.026

McGough, E. L., Lin, S. Y., Belza, B., Becofsky, K. M., Jones, D. L., Liu, M., et al. (2017). A scoping review of physical performance outcome measures used in exercise interventions for older adults with Alzheimer disease and related dementias. *J. Geriatr. Phys. Ther.* 42, 28–74. doi: 10.1519/JPT.0000000000000159

Moraes, H., Ferreira, C., Deslandes, A., Cagy, M., Pompeu, F., Ribeiro, P., et al. (2007). Beta and alpha electroencephalographic activity changes after acute exercise. *Arq. Neuropsiquiatr.* 65, 637–641. doi: 10.1590/s0004-282x2007000400018

Moretti, D. V., Babiloni, C., Binetti, G., Cassetta, E., Forno, G. D., Ferreric, F., et al. (2004). Individual analysis of EEG frequency and band power in mild Alzheimer's disease. *Clin. Neurophysiol.* 115, 299–308. doi: 10.1016/s1388-2457(03)00345-6

Morris, J. K., Vidoni, E. D., Johnson, D. K., Sciver, A. V., Mahnken, J. D., Honea, R. A., et al. (2017). Aerobic exercise for Alzheimer's disease: a randomized controlled pilot trial. *PLoS One* 12:e0170547. doi: 10.1371/journal.pone.0170547

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Ouanes, S., and Popp, J. (2019). High cortisol and the risk of dementia and Alzheimer's disease: a review of the literature. *Front. Aging Neurosci.* 11:43. doi: 10.3389/fnagi.2019.00043

Paillard, T., Rolland, Y., and de Souto Barreto, P. (2015). Protective effects of physical exercise in Alzheimer's disease and Parkinson's disease: a narrative review. *J. Clin. Neurol.* 11, 212–219. doi: 10.3988/jcn.2015.11.3.212

Parvin, E., Mohammadian, F., Amani-Shalamzari, S., Bayati, M., and Tazesh, B. (2020). Dual-task training affect cognitive and physical performances and brain wave ratio of patients with Alzheimer's disease: a randomized controlled trial. *ResearchSquare* [Preprint]. doi: 10.21203/rs.3.rs-64074/v1

Rao, A. K., Chou, A., Bursley, B., Smulofsky, J., and Jezequel, J. (2014). Systematic review of the effects of exercise on activities of daily living in people with Alzheimer's disease. *Am. J. Occup. Ther.* 68, 50–56. doi: 10.5014/ajot.2014.009035

Rikli, R. E., and Jones, C. J. (2013). *Senior Fitness Test Manual.* 2nd Edn. Champaign, IL: Human kinetics.

Roberts, H. C., Denison, H. J., Martin, H. J., Patel, H. P., Syddall, H., Cooper, C., et al. (2011). A review of the measurement of grip strength in clinical and epidemiological studies: towards a standardised approach. *Age Ageing* 40, 423–429. doi: 10.1093/ageing/afr051

Rubin, R. D., Watson, P. D., Duff, M. C., and Cohen, N. J. (2014). The role of the hippocampus in flexible cognition and social behavior. *Front. Hum. Neurosci.* 8:742. doi: 10.3389/fnhum.2014.00742

Santana-Sosa, E., Barriopedro, M. I., López-Mojares, L. M., Pérez, M., and Lucia, A. (2008). Exercise training is beneficial for Alzheimer's patients. *Int. J. Sports Med.* 29, 845–850. doi: 10.1055/s-2008-1038432

Schulz, K. F., Altman, D. G., Moher, D., and CONSORT Group (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC. Med.* 8:18. doi: 10.1186/1741-7015-8-18

Sheppard, J. M., and Triplett, N. T. (2016). "Program design for resistance training," in *Essentials of Strength Training and Conditioning*, ed G. G. Haff and N. T. Triplett (Champaign, IL: Human Kinetics), 439–469.

Sikaroodi, H., Majidi, A., Samadi, S., Shirzad, H., Aghdam, H., Azimi Kia, A., et al. (2012). Evaluating reliability of the montreal cognitive assessment test and its agreement with neurologist diagnosed among patients with cognitive complaints. *Police Med.* 1, 11–17.

Techayusukcharoen, R., Iida, S., and Aoki, C. (2019). Observing brain function *via* functional near-infrared spectroscopy during cognitive program training (dual task) in young people. *J. Phys. Ther. Sci.* 31, 550–555. doi: 10.1589/jpts.31.550

Tolea, M. I., and Galvin, J. E. (2016). The relationship between mobility dysfunction staging and global cognitive performance. *Alzheimer Dis. Assoc. Disord.* 30, 230–236. doi: 10.1097/wad.0000000000000136

Voss, M. W., Heo, S., Prakash, R. S., Erickson, K. I., Alves, H., Chaddock, L., et al. (2013). The influence of aerobic fitness on cerebral white matter integrity and cognitive function in older adults: results of a one-year exercise intervention. *Hum. Brain Mapp.* 34, 2972–2985. doi: 10.1002/hbm.22119

Wang, R., and Holsinger, D. R. M. (2018). Exercise-induced brain-derived neurotrophic factor expression: therapeutic implications for Alzheimer's dementia. *Ageing Res. Rev.* 48, 109–121. doi: 10.1016/j.arr.2018.10.002

Xiao, T., Yang, L., Smith, L., Loprinzi, P. D., Veronese, N., Yao, J., et al. (2020). Correlation between cognition and balance among middle-aged and older adults observed through a tai chi intervention program. *Front. Psychol.* 11:668. doi: 10.3389/fpsyg.2020.00668

Conflict of Interest: The 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.

ORIGINAL RESEARCH
published: 15 January 2021
doi: 10.3389/fpsyg.2020.624137

Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech

R'mani Haulcy and James Glass*

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, United States

Alzheimer's Disease (AD) is a form of dementia that affects the memory, cognition, and motor skills of patients. Extensive research has been done to develop accessible, cost-effective, and non-invasive techniques for the automatic detection of AD. Previous research has shown that speech can be used to distinguish between healthy patients and afflicted patients. In this paper, the ADReSS dataset, a dataset balanced by gender and age, was used to automatically classify AD from spontaneous speech. The performance of five classifiers, as well as a convolutional neural network and long short-term memory network, was compared when trained on audio features (i-vectors and x-vectors) and text features (word vectors, BERT embeddings, LIWC features, and CLAN features). The same audio and text features were used to train five regression models to predict the Mini-Mental State Examination score for each patient, a score that has a maximum value of 30. The top-performing classification models were the support vector machine and random forest classifiers trained on BERT embeddings, which both achieved an accuracy of 85.4% on the test set. The best-performing regression model was the gradient boosting regression model trained on BERT embeddings and CLAN features, which had a root mean squared error of 4.56 on the test set. The performance on both tasks illustrates the feasibility of using speech to classify AD and predict neuropsychological scores.

Keywords: Alzheimer's disease, dementia detection, speech, BERT, i-vectors, x-vectors, word vectors, MMSE prediction

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Johannes Schröder,
Heidelberg University, Germany
Sofia De La Fuente Garcia,
University of Edinburgh,
United Kingdom

****Correspondence:***
R'mani Haulcy
rhaulcy@mit.edu

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *30 October 2020*
Accepted: *09 December 2020*
Published: *15 January 2021*

Citation:
Haulcy R and Glass J (2021) Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech. Front. Psychol. 11:624137. doi: 10.3389/fpsyg.2020.624137

1. INTRODUCTION

Alzheimer's Disease (AD) is a progressive, neurodegenerative disease that affects the lives of more than 5 million Americans every year. The number of Americans living with AD is expected to be more than double that number by 2050. AD is a deadly and costly disease that has negative emotional, mental, and physical implications for those afflicted with the disease and their loved ones (Alzheimer's Association, 2019).

There is currently no cure for AD (Yadav, 2019) and early detection is imperative for effective intervention to occur (De Roeck et al., 2019). Currently, AD is diagnosed using PET imaging and cerebrospinal fluid exams to measure the concentration of amyloid plaques in the brain, a costly and invasive process (Land and Schaffer, 2020). A more cost-effective, non-invasive and easily-accessible technique is needed for detecting AD.

Previous research has shown that speech can be used to distinguish between healthy and AD patients (Pulido et al., 2020). Some researchers have focused on developing new machine learning model architectures to improve detection (Chen et al., 2019; Chien et al., 2019; Liu et al., 2020), while others have used language models (Guo et al., 2019) to classify AD. Others have focused on trying to extract acoustic and text features that capture information indicative of AD. These features include non-verbal features, such as the length of segments and the amount of silence (König et al., 2015). Other researchers have used linguistic and audio features extracted from English speech (Fraser et al., 2016; Gosztolya et al., 2019), as well as Turkish speech (Khodabakhsh et al., 2015). Prosodic features have been extracted from English speech (Ossewaarde et al., 2019; Nagumo et al., 2020; Qiao et al., 2020) and German speech (Weiner et al., 2016) to classify AD, and so have paralinguistic acoustic features (Haider et al., 2019). Other researchers have chosen to focus on the type of speech data that is used instead of the type of model or type of features and have used speech from people performing multiple tasks to improve generalizability (Balagopalan et al., 2018). This provides a brief summary of the work that has been done in the past few years. A more extensive review of the background literature can be found in the review paper of de la Fuente Garcia et al. (2020).

Although promising research has been done, the datasets that have been used are often imbalanced and vary across studies, making it difficult to compare the effectiveness of different modalities. Two recent review papers (Voleti et al., 2019; de la Fuente Garcia et al., 2020) explain that an important future direction for the detection of cognitive impairment is providing a balanced, standardized dataset that will allow researchers to compare the effectiveness of different classification techniques and feature extraction methods. This is what the ADReSS challenge attempted to do. The ADReSS challenge provided an opportunity for different techniques to be performed on a balanced dataset that alleviated the common biases associated with other AD datasets and allowed those techniques to be directly compared.

Previous work has been done using the ADReSS dataset. Some researchers only participated in the AD classification task (Edwards et al., 2020; Pompili et al., 2020; Yuan et al., 2020), others only participated in the Mini-Mental State Examination (MMSE) prediction task (Farzana and Parde, 2020), and others participated in both tasks (Balagopalan et al., 2020; Cummins et al., 2020; Koo et al., 2020; Luz et al., 2020; Martinc and Pollak, 2020; Pappagari et al., 2020; Rohanian et al., 2020; Sarawgi et al., 2020; Searle et al., 2020; Syed et al., 2020). The best performance on the AD classification task was achieved by Yuan et al. (2020), who obtained an accuracy of 89.6% on the test set using linguistic features extracted from the transcripts, as well as encoded pauses. The best performance on the MMSE prediction task was achieved by Koo et al. (2020), who obtained a root mean squared error (RMSE) of 3.747 using a combination of acoustic and textual features.

In this paper, audio features (i-vectors and x-vectors) and text features (word vectors, BERT embeddings, LIWC features, and CLAN features) were extracted from the data and used to train several classifiers, neural networks, and regression models to detect AD and predict MMSE scores. I-vectors and x-vectors, originally intended to be used for speaker verification, have been shown to be effective for detecting AD (López et al., 2019) and other neurodegenerative diseases, such as Parkinson's Disease (Botelho et al., 2020; Moro-Velazquez et al., 2020). Word vectors have also been shown to be useful for detecting AD (Hong et al., 2019). I-vectors, x-vectors, and BERT embeddings have been used with the ADReSS dataset to classify AD (Pompili et al., 2020; Yuan et al., 2020) and predict MMSE scores (Balagopalan et al., 2020). Pompili et al. (2020) used the same audio features that we used and also used BERT embeddings, but they did not apply their techniques to the MMSE prediction task and their best fusion model obtained lower performance on the classification task than our best model. The difference between our work and the work of Balagopalan et al. (2020) and Yuan et al. (2020) is that they finetuned a pre-trained BERT model on the ADReSS data and used that model for classification and regression, whereas we used a pre-trained BERT model as a feature extractor and then trained different classifiers and regressors on the extracted BERT embeddings.

CLAN features were used in the baseline paper (Luz et al., 2020) and were combined with BERT embeddings in this paper to explore whether performance improved. Lastly, LIWC features have been used to distinguish between AD patients and healthy controls in the past (Shibata et al., 2016) but the dataset was very small (nine AD patients and nine healthy controls), and to our knowledge, literature using LIWC for Alzheimer's detection is limited. However, LIWC features have been used to analyze other aspects of mental health (Tausczik and Pennebaker, 2010) and may be useful in the field of AD. For these reasons, we wanted to further explore whether LIWC features could be useful for AD detection and MMSE prediction. Even though our results do not out-perform the best performance on the classification and MMSE prediction tasks, the approaches we employ are different than previous approaches, which provides additional insight into which techniques are best for AD classification and MMSE prediction.

2. MATERIALS AND METHODS

2.1. ADReSS Dataset

The ADReSS challenge dataset consists of audio recordings, transcripts, and metadata (age, gender, and MMSE score) for non-AD and AD patients. The dataset is balanced by age, gender, and number of non-AD vs. AD patients, with there being 78 patients for each class. The audio recordings are of each patient completing the cookie theft picture description task, where each participant describes what they see in the cookie theft image. This task has been used for decades to diagnose and compare AD and non-AD patients (Cooper, 1990; Mendez and Ashla-Mendez, 1991; Giles et al., 1996; Bschor et al., 2001; Mackenzie et al., 2007; Choi, 2009; Hernández-Domínguez et al., 2018; Mueller et al., 2018), as well as patients with other forms of cognitive impairment, and was originally designed as part of an aphasia examination (Goodglass and Kaplan, 1983).

Normalized audio chunks were provided for each speaker, in which a voice activity detection (VAD) system was applied to each patient's recording to split it into several chunks. The VAD system used a log energy threshold value to detect the sections of the audio that contained speech by ignoring sounds

TABLE 1 | Age and gender details for patients in the training set, as well as the average MMSE scores, average years of education, and corresponding standard deviations (sd), for the patients in each age interval.

	AD				Non-AD			
Age interval	**Male**	**Female**	**MMSE (sd)**	**Educ. (sd)**	**Male**	**Female**	**MMSE (sd)**	**Educ. (sd)**
[50, 55)	1	0	30.0 (n/a)	12.0 (n/a)	1	0	29.0 (n/a)	12.0 (n/a)
[55, 60)	5	4	16.3 (4.9)	12.4 (1.7)	5	4	29.0 (1.3)	15.8 (2.8)
[60, 65)	3	6	18.3 (6.1)	12.5 (2.1)	3	6	29.3 (1.3)	13.1 (2.3)
[65, 70)	6	10	16.9 (5.8)	12.8 (2.0)	6	10	29.1 (0.9)	13.8 (3.1)
[70, 75)	6	8	15.8 (4.5)	10.4 (2.6)	6	8	29.1 (0.8)	14.9 (3.4)
[75, 80)	3	2	17.2 (5.4)	10.6 (2.7)	3	2	28.8 (0.4)	14.2 (3.7)
Full set	24	30	17.0 (5.5)	11.9 (2.4)	24	30	29.1 (1.0)	14.3 (3.1)

TABLE 2 | Age and gender details for patients in the test set, as well as the average MMSE scores, average years of education, and corresponding standard deviations (sd), for the patients in each age interval.

	AD				Non-AD			
Age interval	**Male**	**Female**	**MMSE (sd)**	**Educ. (sd)**	**Male**	**Female**	**MMSE (sd)**	**Educ. (sd)**
[50, 55)	1	0	23.0 (n/a)	20.0 (n/a)	1	0	28.0 (n/a)	12.0 (n/a)
[55, 60)	2	2	18.7 (1.0)	12.5 (1.0)	2	2	28.5 (1.2)	13.7 (2.1)
[60, 65)	1	3	14.7 (3.7)	13.2 (2.2)	1	3	28.7 (0.9)	12.2 (0.5)
[65, 70)	3	4	23.2 (4.0)	11.7 (1.9)	3	4	29.4 (0.7)	13.3 (1.4)
[70, 75)	3	3	17.3 (6.9)	12.8 (3.6)	3	3	28.0 (2.4)	13.2 (1.8)
[75, 80)	1	1	21.5 (6.3)	13.0 (1.4)	1	1	30.0 (0.0)	14.0 (2.8)
Full set	11	13	19.5 (5.3)	12.8 (2.7)	11	13	28.8 (1.5)	13.2 (1.6)

below a certain threshold. A 65 dB log energy threshold value was used, along with a maximum duration of 10 s per chunk. Volume normalization involves changing the overall volume of an audio file to reach a certain volume level. There was some variation in the recording environment for each audio file, such as microphone placement, which lead to variation in the volume levels for different recordings. The volume of each chunk was normalized relative to its largest value to remove as much variation from the recordings as possible. Each patient had an average of 25 normalized audio chunks, with a standard deviation of 13 chunks. The CHAT coding system (MacWhinney, 2014) was used to create the transcripts.

The ADReSS dataset is a subset of the Pitt corpus (Becker et al., 1994), which is a dataset that contains 208 patients with possible and probable AD, 104 healthy patients, and 85 patients with an unknown diagnosis. The dataset consists of transcripts and recorded responses from the participants for the cookie theft picture description task, a word fluency task, and a story recall task. In order to provide additional in-domain data for training some of the feature extractors, the cookie theft data for patients not included in the ADReSS dataset was separated from the Pitt corpus and used for pre-training. Normalized audio chunks for this data were created using the steps mentioned above. The pre-training process is described in greater detail in section 2.2.2.

The age and gender distributions, along with the average MMSE scores, average years of education, and corresponding standard deviations, for the training and test sets, can be seen in **Tables 1**, **2**. Education information was not provided with the ADReSS dataset. However, the Pitt corpus did have education information and was cross-referenced with the ADReSS dataset to determine which patients overlapped and to extract each patient's education information. A total of 108 patients (54 non-AD and 54 AD) were selected from the full dataset to create the training set, and the remaining 48 patients (24 non-AD and 24 AD) were used for the test set. For both the training and test sets, an equal number of AD and non-AD patients were included for each age group and the number of male and female AD and non-AD patients was the same for each age group. For the training set, the average MMSE score for the AD patients was 17.0 and the average MMSE score for the non-AD patients was 29.1. The average years of education were 11.9 and 14.3 for the AD and non-AD patients, respectively. For the test set, the AD patients had an average MMSE score of 19.5 and the non-AD patients had an average MMSE score of 28.8. The average years of education were 12.8 and 13.2 for the AD and non-AD patients, respectively.

2.2. Feature Extraction

2.2.1. Text Features: fastText Word Vectors, BERT Embeddings, LIWC, and CLAN Features

FastText is an open-source library that is used to classify text and learn text representations. A fastText model pre-trained on Common Crawl and Wikipedia was used to extract word vectors (Grave et al., 2018) from the transcripts of each speaker. PyLangAcq (Lee et al., 2016), a Python library designed to handle CHAT transcripts, was used to extract the sentences from the CHAT transcript of each participant. A 100-dimensional word vector was computed for each word in each sentence, including punctuation. A dimension of 100 was chosen because this was the value recommended on the fastText website and 100 was compatible with the size of the pre-trained model. The longest sentence had a total of 47 words. For this reason, every sentence was padded to a length of 47, resulting in a (47, 100) representation for each utterance.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) models are text classification models that have achieved state-of-the-art results on a wide variety of natural language processing tasks and they provide high-level language representations called embeddings. Embeddings are vector representations of words or phrases and are useful for representing language because the embeddings often capture information that is universal across different tasks. Keras BERT was used to load an official, pre-trained BERT model and that model was used to extract embeddings of shape (x,768) for each utterance in the transcript of each speaker, where x depends on the length of the input. After embeddings were extracted for each utterance, the largest embedding had an x value of 60. For this reason, the remaining embeddings were padded to be the same shape, resulting in a (60,768) embedding for each utterance. For both the word vectors and the BERT embeddings, features were extracted at the utterance level, resulting in a total of 1,492 embeddings in the training set and 590 embeddings in the test set.

Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010) features were also extracted from the transcripts of each speaker. The LIWC program takes in a transcript and outputs a 93-dimensional vector consisting of word counts for different emotional and psychological categories, such as emotional tone, authenticity, and clout, to name a few. The Computerized Language Analysis (CLAN) program was also used to extract linguistic features from the transcripts of each speaker. The EVAL function was used to extract summary data, including duration, percentage of word errors, number of repetitions, etc. This extraction resulted in a 34-dimensional vector for each speaker. The CLAN features were used as linguistic features in the baseline paper (Luz et al., 2020). In this paper, the CLAN features were combined with the BERT embeddings to explore whether combining the features improved performance. Both the LIWC and CLAN features were extracted at the subject-level, resulting in 108 vectors in the training set and 54 vectors in the test set.

2.2.2. Audio Features: I-Vectors and X-Vectors

VoxCeleb 1 and 2 (Nagrani et al., 2017) are datasets consisting of speech that was extracted from YouTube videos of interviews with celebrities. I-vector and x-vector systems (Snyder et al., 2017, 2018) pre-trained on VoxCeleb 1 and 2 were used to extract i-vectors and x-vectors from the challenge data. The i-vector and x-vector systems were built using Kaldi (Povey et al., 2011), which is a toolkit that is used for speech recognition. The pre-trained VoxCeleb models were also used to train additional extractors using the original Kaldi recipes. The original VoxCeleb models were used to initialize the i-vector and x-vector extractors and then those extractors were trained on the remaining in-domain Pitt data. I-vector and x-vector extractors were also trained on only the in-domain Pitt data to explore whether a small amount of in-domain data is better for performance than a large amount of out-of-domain data. For each type of extractor, the normalized audio chunks provided with the challenge dataset were first resampled with a sampling rate of 16kHz, a single channel, and 16 bits, to match the configuration of the VoxCeleb data. The Kaldi toolkit was then used to extract the Mel-frequency cepstral coefficients (MFCCs), compute the voice activation detection (VAD) decision, and extract the i-vectors and x-vectors. The x-vectors had a length of 512, while the i-vectors had a length of 400. There were a total of 2,834 i-vectors and 2,834 x-vectors, one i-vector and x-vector for each normalized audio chunk.

2.3. Experimental Approach

2.3.1. Classifiers

Five classifiers were trained on the text and audio features explained in sections 2.2.1 and 2.2.2: linear discriminant analysis (LDA), the decision tree (DT) classifier, the k-nearest neighbors classifier with the number of neighbors set to 1 (1NN), a support vector machine (SVM) with a linear kernel and regularization parameter set to 0.1, and a random forest (RF) classifier. The classifiers were implemented in Python using the scikit-learn library. The word vectors and BERT embeddings were averaged before being used to train the scikit-learn classifiers, resulting in utterances represented by 100-dimensional vectors and 768-dimensional vectors, respectively. When the LIWC and CLAN features were combined with the averaged BERT embeddings, the subject-level LIWC/CLAN vector was concatenated with each utterance-level BERT embedding belonging to that subject. Standard scaling is commonly applied to data before using machine learning estimators to avoid the poor performance that is sometimes seen when the features are not normally distributed (i.e., Gaussian with a mean of 0 and unit variance). Because we were combining different types of features with different data distributions, standard scaling was applied to the features after the LIWC/CLAN vectors were concatenated with the BERT embeddings so that the data would be normally distributed before training and testing.

2.3.2. Regressors

Five regression models were also trained on the text and audio features explained in sections 2.2.1 and 2.2.2 for the MMSE prediction task: linear regression (LR), decision tree (DT) regressor, k-nearest neighbor regressor with the number of neighbors set to 1 (1NN), support vector machine (SVM), and a gradient-boosting regressor (grad-boost). The regression models were implemented in Python using the scikit-learn library. Just as with the classifiers, the word vectors and BERT embeddings were averaged before being used to train the scikit-learn regressors. When the LIWC and CLAN features were combined with the BERT embeddings, the subject-level LIWC/CLAN vector was concatenated with each utterance-level BERT embedding belonging to that subject, and after the features were concatenated, standard scaling was applied.

2.3.3. Dimensionality Reduction

The classifiers and regressors mentioned in sections 2.3.1 and 2.3.2 were trained with different dimensionality reduction techniques to see if applying dimensionality reduction improves performance. Feature sets were created with no dimensionality reduction, with LDA, and with principal component analysis (PCA), and each classifier was trained on each feature set to see what effect dimensionality reduction had on performance. The dimensionality reduction techniques were applied to all of the audio and text features. When LDA was applied, the features were reduced to 1 dimension for the classification task and 23 dimensions for the regression task. With PCA, different dimension values were selected manually. The best results and corresponding dimension values can be seen in the Results section.

2.3.4. Neural Networks

A bidirectional long short-term memory (LSTM) network and a convolutional neural network (CNN) were also trained on the word vectors to see if the neural networks could extract some temporal information that would lead to better performance compared to the classifiers mentioned in section 2.3.1. The topologies of the two networks are shown in **Figure 1**. The LSTM model had one bidirectional LSTM layer with eight units, a dropout rate of 0.2, and a recurrent dropout rate of 0.2. The CNN model had the following layers: three 2D convolution layers with 32, 64, and 128 filters, respectively, rectified linear unit (ReLu) activation and a kernel size of 3, one 2D max pooling layer with a pool size of 3, one dropout layer with a rate of 0.5, and one 2D global max pooling layer. For both models, the output was passed into a dense layer with sigmoid activation. Both models were implemented in Python using Keras and were trained with an Adam optimizer. The CNN was trained with a learning rate of 0.001, and the LSTM was trained with a learning rate of 0.01.

3. RESULTS

3.1. Classification

3.1.1. Cross-Validation

In order to stay consistent with the baseline paper, each of the classifiers and neural networks were evaluated on the challenge training set using leave-one-subject-out (LOSO) cross-validation, where there was no speaker overlap between the training and test sets for each split. Each model was trained and tested at the utterance level, where each utterance was classified as belonging to a patient with or without AD. Then majority vote (MV) classification was used to assign a label to each speaker based on the label that was assigned most to the speaker's utterances.

The MV classification accuracy (the number of correctly classified speakers divided by the total number of speakers), for each feature type can be seen in **Table 3**. The accuracies are presented as decimals and are rounded to 3 decimal places to match the form of the accuracies in the baseline paper. For all of the features, the LDA classifier trained on LDA-reduced features performed the same as the LDA classifier trained on features with no dimensionality reduction. Although there was no difference in performance, results are included for completeness.

The LSTM model trained on word vectors had an average accuracy of **0.787**, while the CNN model had an average accuracy of **0.704**. The highest-performing classifier trained on text features was the SVM classifier trained on a combination of BERT embeddings and CLAN features with PCA dimensionality reduction applied, which had an average accuracy of 0.898. The highest-performing classifier trained on audio features was the LDA classifier trained on x-vectors that were extracted using a system that was pre-trained on VoxCeleb and in-domain Pitt data. PCA dimensionality reduction was applied and the classifier had an average accuracy of 0.657.

The highest-performing classifiers for each feature type, except for the classifiers trained on x-vectors that were extracted from a system trained on just Pitt data, performed better than the highest-performing audio and text baseline classifiers that were evaluated using LOSO on the training set, which had an average accuracy of 0.565 and 0.768, respectively (Luz et al., 2020).

3.1.2. Held-Out Test Set

The MV classification accuracies on the test set for each of the classifiers can be seen in **Table 4**. The highest-performing text classifiers were the SVM classifier with no dimensionality reduction and the RF classifier with PCA dimensionality reduction, both trained on BERT embeddings. Both classifiers had an average accuracy of 0.854. The highest-performing audio classifier was the 1NN classifier trained on i-vectors that were extracted using systems pre-trained on VoxCeleb with PCA dimensionality reduction applied, which had an average accuracy of 0.563.

The highest-performing text classifiers outperformed the baseline text classifier, which was an LDA classifier trained on CLAN features with an average accuracy of 0.75. The highest-performing audio classifiers did not outperform the baseline audio classifier, which was an LDA classifier trained on ComParE openSMILE features with an average accuracy of 0.625.

3.2. MMSE Prediction

3.2.1. Cross-Validation

For the MMSE prediction task, one of the speakers in the training set did not have an MMSE score and was excluded from training. Each of the regressors was evaluated on the challenge training set using LOSO cross-validation, where there was no speaker overlap between the training and test sets for each split. Each model was trained and tested at the utterance level, where an MMSE score was predicted for each utterance. Then the predicted MMSE scores of the utterances belonging to a patient were averaged to assign one MMSE score to that patient. Lastly, the RMSE between the predicted and ground truth MMSE scores was computed.

The average RMSE scores for each feature type can be seen in **Table 5**. For all of the features, the LR regressor trained on LDA-reduced features performed the same as the LR regressor trained on features with no dimensionality reduction. Although there was no difference in performance, results are included for completeness.

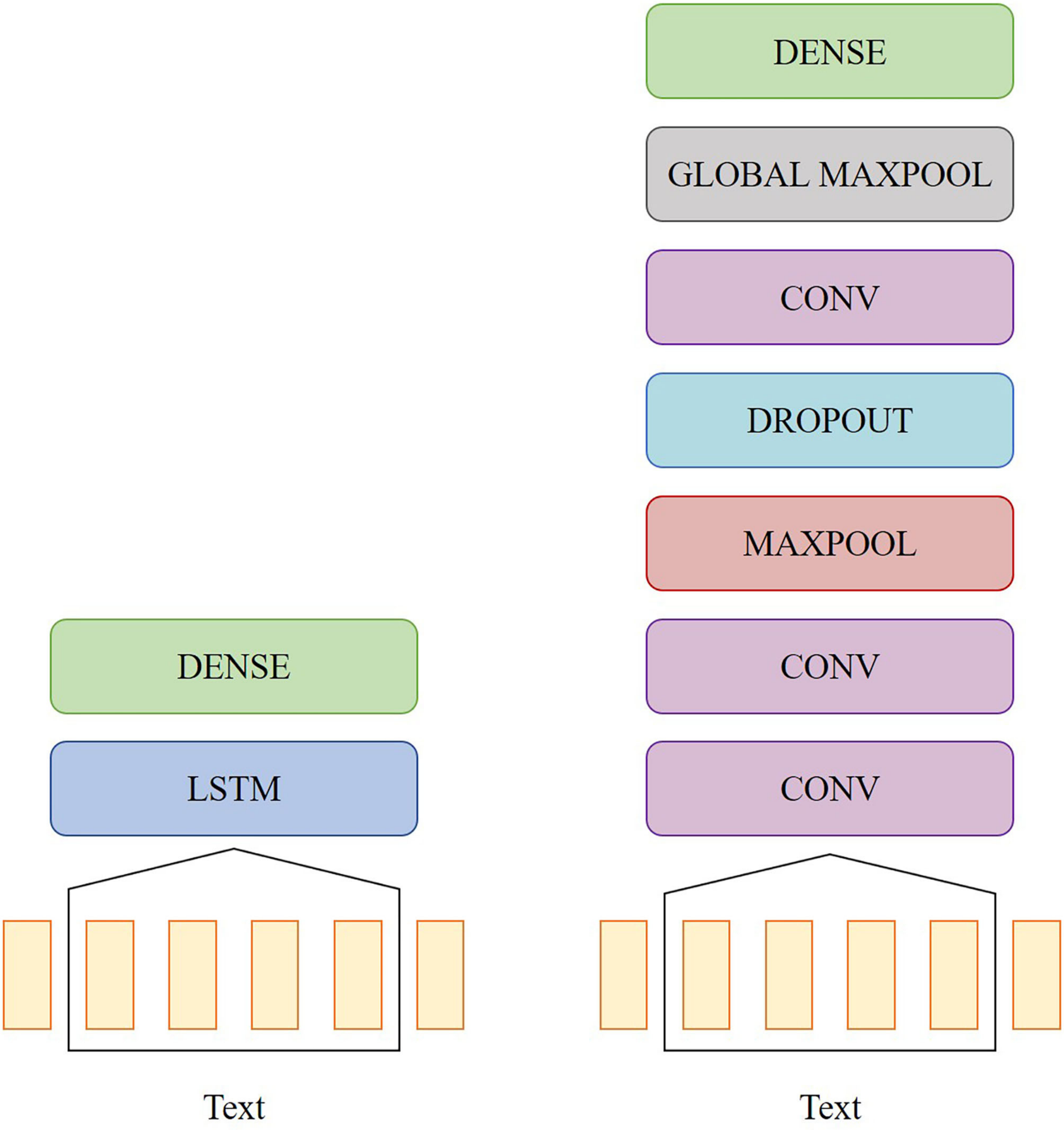


FIGURE 1 | Diagrams of the network topology for the LSTM model (left) and the CNN model (right).

The best-performing regressor trained on text features was the LR regressor trained on BERT embeddings combined with LIWC and CLAN features with PCA dimensionality reduction applied, which had an RMSE score of 3.774. The best-performing regressor trained on audio features was the DT regressor trained on x-vectors that were extracted using a system pre-trained on Pitt. LDA dimensionality reduction was applied and the RMSE score was 6.073.

The best-performing text regressors for every feature type, except for BERT embeddings and word vectors, performed better than the baseline text regressor that was evaluated using LOSO on the training set, which had an RMSE score of 4.38. The best-performing audio regressors for every feature type

TABLE 3 | LOSO accuracies for each of the classifiers. The best-performing models for each feature type are red.

Features	Dim. Red. (n_comp)	LDA	DT	1NN	SVM	RF
LIWC	None	0.741	0.593	0.620	**0.833**	0.778
	LDA (1)	0.741	0.750	0.750	0.731	0.750
	PCA (20)	0.778	0.620	0.704	0.787	0.759
BERT	None	0.713	0.676	0.787	**0.796**	0.769
	LDA (1)	0.713	0.657	0.667	0.713	0.657
	PCA (2)	0.630	0.648	0.602	0.546	0.694
	PCA (20)	0.750	0.713	0.722	0.769	**0.796**
BERT + LIWC	None	0.750	0.657	0.667	**0.824**	0.806
	LDA (1)	0.750	0.731	0.731	0.741	0.731
	PCA (20)	**0.824**	0.620	0.657	**0.824**	0.796
BERT + CLAN	None	0.778	0.657	0.759	0.824	0.750
	LDA (1)	0.778	0.769	0.769	0.787	0.769
	PCA (20)	0.824	0.630	0.657	**0.898**	0.778
BERT + LIWC + CLAN	None	0.593	0.731	0.713	0.815	0.806
	LDA (1)	0.593	0.611	0.611	0.593	0.611
	PCA (20)	**0.833**	0.731	0.713	0.815	0.787
word vectors	None	0.759	0.731	0.694	0.259	0.694
	LDA (1)	0.759	0.741	0.731	0.759	0.759
	PCA (2)	0.676	0.620	0.565	0.259	0.620
	PCA (70)	**0.796**	0.648	0.759	**0.796**	0.787
i-vectors (VoxCeleb)	None	0.574	0.423	0.454	0.574	0.500
	LDA (1)	0.574	0.500	0.500	0.574	0.500
	PCA (2)	0.491	0.500	**0.602**	0.519	0.491
	PCA (10)	0.528	0.556	0.546	0.491	0.528
i-vectors (Pitt)	None	0.528	0.491	0.500	0.509	**0.593**
	LDA (1)	0.528	0.537	0.537	0.537	0.537
	PCA (2)	0.463	0.500	0.528	0.343	0.546
	PCA (20)	0.565	0.537	0.528	0.565	0.565
i-vectors (VoxCeleb + Pitt)	None	0.528	0.509	0.500	0.528	0.556
	LDA (1)	0.528	0.519	0.519	0.528	0.519
	PCA (20)	0.519	0.528	0.574	0.472	**0.620**
x-vectors (VoxCeleb)	None	0.583	0.620	0.509	0.546	0.574
	LDA (1)	0.583	0.593	0.593	0.583	0.593
	PCA (2)	0.472	0.537	0.491	0.454	0.491
	PCA (40)	**0.639**	0.583	0.528	**0.639**	0.583
x-vectors (Pitt)	None	**0.546**	**0.546**	0.472	0.528	0.481
	LDA (1)	**0.546**	0.500	0.500	0.537	0.500
	PCA (40)	0.537	0.481	0.435	0.528	0.491
x-vectors (VoxCeleb + Pitt)	None	0.639	0.602	0.519	0.620	0.509
	LDA (1)	0.639	0.509	0.509	0.630	0.509
	PCA (40)	**0.657**	0.574	0.546	0.593	0.593

TABLE 4 | Accuracies for classifiers evaluated on the test set. The test set results for the best-performing models during cross-validation are red.

Features	Dim. Red. (n_comp)	LDA	DT	1NN	SVM	RF
LIWC	None	0.583	0.708	0.583	**0.688**	0.812
	LDA (1)	0.583	0.583	0.583	0.583	0.583
	PCA (20)	0.771	0.646	0.583	0.792	0.667
BERT	None	0.604	0.708	0.771	**0.854**	0.750
	LDA (1)	0.604	0.604	0.646	0.604	0.604
	PCA (2)	0.688	0.562	0.542	0.729	0.625
	PCA (20)	0.833	0.646	0.750	0.812	**0.854**
BERT + LIWC	None	0.583	0.667	0.688	**0.729**	0.812
	LDA (1)	0.583	0.583	0.583	0.583	0.583
	PCA (20)	**0.792**	0.708	0.771	**0.771**	0.792
BERT + CLAN	None	0.729	0.750	0.771	0.812	0.812
	LDA (1)	0.729	0.708	0.708	0.708	0.708
	PCA (20)	0.729	0.708	0.667	**0.771**	0.792
BERT + LIWC + CLAN	None	0.625	0.688	0.750	0.750	0.812
	LDA (1)	0.625	0.667	0.667	0.625	0.667
	PCA (20)	**0.812**	0.604	0.729	0.812	0.812
word vectors	None	0.813	0.688	0.667	0.500	0.833
	LDA (1)	0.813	0.750	0.771	0.813	0.750
	PCA (2)	0.729	0.542	0.500	0.500	0.667
	PCA (70)	**0.812**	0.562	0.688	**0.500**	0.771
i-vectors (VoxCeleb)	None	0.542	0.563	0.521	0.625	0.625
	LDA (1)	0.542	0.521	0.521	0.542	0.521
	PCA (2)	0.750	0.625	**0.563**	0.708	0.729
	PCA (10)	0.562	0.542	0.438	0.583	0.562
i-vectors (Pitt)	None	0.417	0.521	0.521	0.438	**0.542**
	LDA (1)	0.417	0.542	0.542	0.417	0.542
	PCA (2)	0.667	0.583	0.708	0.604	0.646
	PCA (20)	0.583	0.542	0.583	0.521	0.479
i-vectors (VoxCeleb + Pitt)	None	0.458	0.521	0.500	0.500	0.563
	LDA (1)	0.458	0.542	0.542	0.458	0.542
	PCA (20)	0.458	0.563	0.604	0.458	**0.479**
x-vectors (VoxCeleb)	None	0.604	0.500	0.500	0.563	0.521
	LDA (1)	0.604	0.604	0.604	0.604	0.604
	PCA (2)	0.625	0.563	0.563	0.625	0.542
	PCA (40)	**0.479**	0.417	0.562	**0.458**	0.479
x-vectors (Pitt)	None	**0.500**	**0.479**	0.417	0.563	0.583
	LDA (1)	**0.500**	0.542	0.542	0.500	0.542
	PCA (40)	0.521	0.563	0.521	0.458	0.542
x-vectors (VoxCeleb + Pitt)	None	0.563	0.604	0.479	0.521	0.583
	LDA (1)	0.563	0.521	0.521	0.563	0.521
	PCA (40)	**0.500**	0.458	0.646	0.479	0.563

performed better than the baseline audio regressor that was evaluated using LOSO on the training set, which had an RMSE score of 7.28.

3.2.2. Held-Out Test Set

The RMSE scores on the test set for each of the regressors can be seen in **Table 6**. The best-performing text regressor was the grad-boost regressor trained on BERT embeddings combined with CLAN features with PCA dimensionality reduction applied, which had an RMSE score of 4.560. The best-performing audio regressor was the 1NN regressor trained on i-vectors extracted using a system pre-trained on VoxCeleb and Pitt with LDA dimensionality reduction applied, which had an RMSE score of 5.694.

TABLE 5 | LOSO RMSE scores for each of the classifiers. The results for the best-performing models for each feature type are red.

Features	Dim. Red. (n_comp)	LR	DT	1NN	SVM	GradBoost
LIWC	None	10.067	5.766	5.626	6.083	**4.014**
	LDA (23)	8.928	8.738	5.224	6.195	7.654
	PCA (20)	4.436	5.383	5.364	6.057	4.640
BERT	None	5.111	5.984	**4.953**	6.111	5.407
	LDA (23)	5.111	6.571	5.805	6.275	6.701
	PCA (2)	6.304	5.628	5.851	6.187	6.034
BERT + LIWC	None	9.475	4.956	4.752	5.919	**4.050**
	LDA (23)	8.515	8.038	5.285	6.821	7.234
	PCA (20)	4.574	5.228	5.680	5.165	4.509
BERT + CLAN	None	4.810	6.265	4.728	6.009	4.100
	LDA (23)	4.810	5.700	4.988	6.173	5.447
	PCA (20)	3.991	5.459	4.842	5.254	**3.969**
BERT + LIWC + CLAN	None	13.877	5.533	4.420	5.846	4.190
	LDA (23)	5.243	5.398	5.482	6.477	5.031
	PCA (20)	**3.774**	5.701	5.023	4.966	4.201
word vectors	None	5.294	5.467	5.204	6.146	5.684
	LDA (23)	5.294	5.158	**4.967**	5.936	5.228
	PCA (2)	6.359	6.061	5.958	6.148	6.241
	PCA (70)	5.419	5.561	4.981	6.177	5.516
i-vectors (VoxCeleb)	None	6.323	6.477	6.612	6.444	6.461
	LDA (23)	6.323	6.366	6.384	6.279	6.443
	PCA (2)	6.576	6.431	6.361	6.290	6.421
	PCA (10)	6.412	6.507	6.524	6.265	**6.264**
i-vectors (Pitt)	None	6.545	6.850	6.239	6.281	6.513
	LDA (23)	6.545	6.524	6.307	6.244	6.499
	PCA (2)	6.624	6.606	6.484	6.323	6.598
	PCA (20)	6.523	6.575	6.577	**6.207**	6.511
i-vectors (VoxCeleb + Pitt)	None	6.298	6.363	6.545	6.243	6.445
	LDA (23)	6.298	6.399	**6.110**	6.231	6.459
	PCA (20)	6.502	6.558	6.655	6.256	6.475
x-vectors (VoxCeleb)	None	6.424	6.400	6.208	6.400	6.369
	LDA (23)	6.424	6.478	6.493	**6.162**	6.413
	PCA (2)	6.618	6.767	6.531	6.381	6.634
	PCA (40)	6.246	6.320	6.517	6.329	6.378
x-vectors (Pitt)	None	6.310	6.534	6.445	6.405	6.504
	LDA (23)	6.310	**6.073**	6.403	6.245	6.318
	PCA (40)	6.471	6.456	6.181	6.369	6.474
x-vectors (VoxCeleb + Pitt)	None	6.385	6.268	6.394	6.401	6.386
	LDA (23)	6.385	6.379	6.230	**6.170**	6.442
	PCA (40)	6.296	6.433	6.411	6.288	6.467

TABLE 6 | RMSE scores for classifiers evaluated on the test set. The results for the best-performing models during cross-validation are red.

Features	Dim. Red. (n_comp)	LR	DT	1NN	SVM	GradBoost
LIWC	None	36.974	7.303	6.403	6.465	**4.862**
	LDA (23)	12.286	9.657	7.388	6.313	8.365
	PCA (20)	4.422	5.967	5.990	6.431	4.383
BERT	None	5.365	5.640	**4.923**	6.169	4.883
	LDA (23)	5.365	7.515	6.017	6.253	7.373
	PCA (2)	5.661	5.858	6.287	6.067	5.691
BERT + LIWC	None	34.420	7.127	5.021	6.103	**5.037**
	LDA (23)	14.905	8.624	5.742	7.189	6.561
	PCA (20)	4.872	7.078	5.159	4.895	4.404
BERT + CLAN	None	4.991	7.218	4.515	6.097	4.901
	LDA (23)	4.991	6.523	5.600	6.422	6.660
	PCA (20)	4.764	7.577	6.413	5.218	**4.560**
BERT + LIWC + CLAN	None	15.465	6.112	4.811	6.023	4.724
	LDA (23)	8.110	6.500	5.753	6.887	6.021
	PCA (20)	**4.800**	6.196	5.532	4.794	5.087
word vectors	None	4.714	5.280	5.129	6.147	5.361
	LDA (23)	4.714	5.111	**5.344**	6.063	4.955
	PCA (2)	5.732	6.452	5.992	6.129	5.803
	PCA (70)	4.785	5.700	5.237	6.169	5.271
i-vectors (VoxCeleb)	None	6.600	6.305	6.269	6.161	6.396
	LDA (23)	6.600	7.056	6.360	6.461	6.820
	PCA (2)	6.194	6.514	6.546	5.999	6.237
	PCA (10)	6.335	6.840	6.298	6.110	**6.386**
i-vectors (Pitt)	None	6.530	6.622	6.758	6.142	6.170
	LDA (23)	6.530	6.712	6.133	5.956	6.473
	PCA (2)	6.225	6.827	6.370	6.151	6.342
	PCA (20)	6.257	6.278	6.110	**6.199**	6.252
i-vectors (VoxCeleb + Pitt)	None	6.292	6.042	7.391	6.158	6.145
	LDA (23)	6.292	6.567	**5.694**	5.905	6.407
	PCA (20)	6.316	6.439	6.607	6.168	6.431
x-vectors (VoxCeleb)	None	6.559	6.665	6.401	6.094	6.309
	LDA (23)	6.559	6.289	6.261	**6.085**	6.312
	PCA (2)	6.167	6.669	6.566	6.089	6.164
	PCA (40)	6.358	6.058	6.189	6.115	6.160
x-vectors (Pitt)	None	6.428	6.483	6.563	6.287	6.333
	LDA (23)	6.428	**6.462**	6.314	6.097	6.423
	PCA (40)	6.424	6.506	6.499	6.322	6.370
x-vectors (VoxCeleb + Pitt)	None	6.644	6.622	6.338	6.096	6.208
	LDA (23)	6.644	6.450	6.188	**6.059**	6.466
	PCA (40)	6.173	6.640	6.488	6.123	6.204

The highest-performing text regressor outperformed the baseline text regressor, which was a DT regressor trained on CLAN features with an RMSE score of 5.20. The highest-performing audio regressor outperformed the baseline audio regressor, which was a DT regressor trained on Multi-resolution Cochleagram (MRCG) openSMILE features that had an RMSE score of 6.14.

3.3. Effects of Education and the Severity of Cognitive Impairment

In order to explore what effect the severity of cognitive impairment and education level had on the classification and MMSE prediction results, the best-performing text and audio models from both tasks were evaluated on smaller subsets of the test set that were split based on education level and MMSE score.

TABLE 7 | Test set accuracies and RMSE scores for different levels of cognitive deficiency and education.

		Text			Audio	
		Classification		MMSE prediction	Classification	MMSE prediction
	Group (num. patients)	SVM	RF	GradBoost	1NN	1NN
MMSE	Healthy (28)	0.857	0.714	3.234	0.500	4.679
	Mild Dementia (8)	0.750	0.750	3.777	0.625	1.801
	Moderate Dementia (8)	0.875	0.625	4.563	0.500	6.224
	Severe Dementia (4)	1.000	0.500	10.241	0.750	12.323
Education	<12 years (5)	0.800	0.600	7.448	1.000	9.329
	12 years (24)	0.792	0.833	4.128	0.458	5.080
	>12 years (19)	0.947	0.684	3.885	0.474	5.138

According to the Alzheimer's Association (2020), an MMSE score of 20–24 corresponds to mild dementia, 13–20 corresponds to moderate dementia, and a score <12 is severe dementia. This information was used to create 4 groups of cognitive severity: healthy (MMSE score ≥25), mild dementia (MMSE score of 20–24), moderate dementia (MMSE score of 13–19), and severe dementia (MMSE score ≤12). The ranges set by the Alzheimer's Association were slightly modified to have unique boundary values.

For education level, the majority of patients had 12 years of education (likely equivalent to completing high school). Because the test set is small, we wanted to limit our experiments to a small number of groups. For the reasons previously mentioned, one education group was for patients that had 12 years of education, another group was for patients with <12 years of education, and the last group included patients that had more than 12 years of education.

The text and audio models were trained on the full training set and then evaluated on each MMSE and education group separately by only testing on patients in the test set that belonged to a particular group. The classification and MMSE prediction results can be seen in **Table 7**. For the MMSE groups, the results showed that the best classification accuracy achieved using a text model was 1.000 and that accuracy was achieved when the SVM classifier was evaluated on patients with severe dementia. The best RMSE achieved using a text model was 3.234 and that RMSE was achieved when the GradBoost regressor was evaluated on healthy patients. For the audio models, the best classification accuracy was 0.750 and was achieved when the 1NN classifier was evaluated on patients with severe dementia. The best RMSE was 1.801 and was achieved when the 1NN was evaluated on patients with mild dementia.

For the education groups, the best classification accuracy achieved using a text model was 0.947, when the SVM classifier was evaluated on patients with more than 12 years of education. The best RMSE was 3.885 and was achieved when the GradBoost model was evaluated on patients with >12 years of education. For the audio models, the best classification accuracy is 1.000 and was achieved when the 1NN was evaluated on patients with <12 years of education. The best RMSE was 5.080 and was achieved when the 1NN was evaluated on patients with 12 years of education.

4. DISCUSSION

The held-out test set results for both tasks show that text classifiers trained on BERT embeddings and text regressors trained on BERT embeddings combined with CLAN features perform better than text classifiers/regressors trained on only CLAN features (baseline text feature set). The results also show that audio classifiers trained on x-vectors and i-vectors, extracted using systems that were pre-trained on VoxCeleb and Pitt data, do not perform better than audio classifiers trained on ComParE openSMILE features (baseline audio feature set). However, audio regressors trained on x-vectors and i-vectors do perform better than audio regressors trained on MRCG openSMILE features when (1) the x-vectors are trained on only out-of-domain data or a combination of in-domain data and out-of-domain data and (2) when i-vectors are trained on a combination of in-domain and out-of-domain data.

We also note that we achieved better test set results on the classification task and equal test set results on the MMSE prediction task using a pre-trained BERT model as a feature extractor as opposed to using BERT as a classifier and regressor as Balagopalan et al. (2020) did. We received classification test set results equal to the BERT results of Yuan et al. (2020), who also used a BERT model as a classifier and added encoded pauses to their training regime. Our results show that BERT embeddings can be used to achieve the BERT model performance of Yuan et al. (2020) without using the BERT model itself as a classifier and without using pause information. However, the results of Yuan et al. (2020) suggest that we could achieve even greater performance if we include pause information in our feature set.

4.1. I-Vector and X-Vector Systems

One possible explanation for the poor performance of the i-vectors and x-vectors on the classification task is the domain-mismatch between the VoxCeleb datasets and the ADReSS dataset. While the pre-trained model may have learned some general representations of speech from the VoxCeleb datasets, it is possible that the type of representations that the model learned were not helpful for distinguishing between the speech of AD and non-AD patients. The VoxCeleb dataset consists of speech extracted from YouTube videos of celebrities being interviewed.

While there is variety in the age, race, and accent of the speakers in the VoxCeleb dataset, which may help improve the ability of a model to distinguish between speakers that differ in these qualities, the nature of the recordings (i.e., background noise, overlapping speech, etc.) varies significantly from the recording environment of the ADReSS data. There is also less variety in the types of speakers present in the ADReSS dataset: they are all within a certain age range and do not seem to have significantly different accents. Therefore, the benefits of the VoxCeleb datasets are not likely to help with the AD classification task and the difference in recording environments likely intensifies the domain-mismatch problem, leading to lower performance. It is possible that i-vectors and x-vectors pre-trained on a different dataset with less of a domain-mismatch would perform better.

The i-vectors extracted from a system that was only trained on Pitt data did not improve performance on the classification task compared to the i-vectors extracted from a system that was trained on VoxCeleb but did improve performance on the MMSE prediction task. Conversely, the x-vectors extracted from a system that was only trained on Pitt did improve performance on the classification task but did not improve performance on the MMSE prediction task. The i-vector and x-vector extractors that we pre-trained on a combination of VoxCeleb and Pitt data led to an improvement in performance on the MMSE prediction task, compared to the performance for i-vectors and x-vectors extracted from a system trained on VoxCeleb. The x-vector performance also improved on the classification task. This shows that a small amount of in-domain data can improve i-vector and x-vector performance for the MMSE prediction task. When choosing between training i-vector and x-vector extractors on a large amount of out-of-domain data, a small amount of in-domain data, or a combination of both, the results suggest that it is best to train on a combination of both.

4.2. Pros and Cons of Linguistic Features

The highest-performing models for both tasks were trained on linguistic features (BERT embeddings). One benefit of using linguistic features is that punctuation can be included. This allows the model to use semantic and syntactical information, such as how often speakers are asking questions ("?" present in the transcript). Also, because the BERT model was pre-trained on BooksCorpus and English Wikipedia, the data that the pre-trained model saw was likely much more general than the VoxCeleb data and using text data meant that the model did not face the issue of the recording-environment mismatch.

However, there are some disadvantages associated with linguistic features. As discussed in the review paper of de la Fuente Garcia et al. (2020), transcript-free approaches to AD detection are better for generalizability and for protecting the privacy of the participants. In order to use linguistic features, the speech must be transcribed, meaning that linguistic features are worse for model generalizability and patient privacy. Using linguistic features depends on the use of automatic speech recognition (ASR) methods, which often have a low level of accuracy, or transcription methods, which can be costly and time-consuming.

Some linguistic features are also content- and language-dependent. There are linguistic features that are not content-dependent, such as word frequency measures, but it is difficult to automate the extraction of content-independent linguistic features (de la Fuente Garcia et al., 2020). For these reasons, it is important that future research explore using AD classification techniques that only require acoustic features.

4.3. Dimensionality Reduction

For the classification task, none of the highest-performing models had LDA dimensionality applied to the feature sets before training. As previously mentioned, the features were reduced to one dimension when LDA was applied. The results suggest that this dimensionality reduction was too extreme for the classification task and did not allow for enough information to be retained in the feature set. Conversely, the majority of the highest-performing classifiers had PCA dimensionality reduction applied to the feature sets before training. This suggests that applying PCA dimensionality reduction to the features before training can be useful for AD classification.

For the MMSE prediction task, the features were reduced to 23 dimensions when LDA was applied. Because the dimension was larger, LDA was more useful for this task. The best-performing audio model had LDA dimensionality reduction applied. PCA dimensionality reduction was also applied for some of the best-performing models, including the top-performing text model. This suggests that applying LDA and PCA dimensionality reduction to the features before training can be useful for MMSE prediction.

4.4. Group Evaluation

The evaluation results for different MMSE and education groups showed that certain MMSE groups can be classified more accurately (healthy, moderate dementia, and severe dementia) while others (mild dementia) are more difficult to classify. This seems very reasonable, as it is expected that more severe forms of dementia would be more easily distinguishable from healthy patients. Also, MMSE scores are predicted least accurately when evaluated on patients with severe dementia, regardless of the type of features used (text or audio).

The education results for the best-performing text-based model showed that patients with more than 12 years of education can be classified with high accuracy (0.947), while patients with exactly 12 years (0.792) and <12 years (0.800) of education are more difficult to classify and are classified with similar accuracy. The MMSE scores of patients with >12 years of education were predicted with the most accuracy.

These results provide some insight into which types of features are best for different levels of dementia and education for the classification and MMSE prediction tasks. However, it is important to note that the evaluation set is small, with as little as four speakers in certain groups (severe dementia). Therefore, these findings may not translate well to larger datasets.

4.5. Conclusions

In this paper, audio and text-based representations of speech were extracted from the ADReSS dataset for the AD classification and MMSE prediction tasks. Different dimensionality reduction techniques were applied to the data before training and testing the classification and regression models to explore whether applying dimensionality reduction techniques improved performance on those tasks. LOSO cross-validation was used to evaluate each of the classifiers and regressors and the models were also evaluated on a held-out test set.

The best-performing text models in this paper outperform the baseline text models on both tasks and the best-performing audio models outperform the baseline on the MMSE prediction task. The audio results suggest that, given access to a large amount of out-of-domain data and a small amount of in-domain data, it is best to use a combination of both to train i-vector and x-vector extractors. The comparison of the dimensionality reduction techniques shows that applying PCA dimensionality reduction to the features before training a classifier can be helpful for this particular AD classification task and possibly for other similar health-related classification tasks. Also, applying LDA and PCA dimensionality reduction to the features before training a regressor can be helpful for MMSE prediction tasks. Lastly, the evaluation results on different MMSE and education groups show that patients with more severe forms of dementia (moderate and severe) and healthy patients are easier to classify than patients with mild dementia, whereas the MMSE scores of severe dementia patients are the most difficult to predict. Patients with more than 12 years of education are the easiest to classify and the MMSE scores of patients with >12 years of education are the easiest to predict.

For future work, it would be interesting to repeat the experiments, particularly the evaluation of audio and text models on MMSE and education groups, on a larger dataset to see whether the findings translate. Another interesting future direction would be relating our findings to apathetic symptoms. Previous research has shown that patients with moderate or severe forms of AD tend to be apathetic (Lueken et al., 2007). Signs of apathy include slow speech, long pauses, and changes in facial expressions (Seidl et al., 2012). These characteristics can be measured using standardized ratings and we can explore whether our findings are consistent with the findings related to other forms of cognitive decline that affect speech.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: in order to gain access to the datasets used in the paper, researchers must become a member of DementiaBank. Requests to access these datasets should be directed to https://dementia.talkbank.org/.

AUTHOR CONTRIBUTIONS

R'mH contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the manuscript. JG contributed to the design of the research and supervised the findings of this work. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported in part by a Mathworks Fellowship and by Millennium Pharmaceuticals, Inc. (a subsidiary of Takeda Pharmaceuticals).

The authors declare that this study received funding from Millenium Pharmaceuticals, Inc. (a subsidiary of Takeda Pharmaceuticals). The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

REFERENCES

Alzheimer's Association (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers Dement*. 15, 321–387. doi: 10.1016/j.jalz.2019.01.010

Alzheimer's Association (2020). *Medical Tests. Alzheimer's Disease and Dementia*. Available online at: https://www.alz.org/alzheimers-dementia/diagnosis/medical_tests (accessed December 3, 2020).

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). "To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection," in *Proceedings of Interspeech 2020* (Shanghai), 2167–2171. doi: 10.21437/Interspeech.2020-2557

Balagopalan, A., Novikova, J., Rudzicz, F., and Ghassemi, M. (2018). The effect of heterogeneous data for Alzheimer's disease detection from speech. *arXiv* 1811.12254.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archiv. Neurol.* 51, 585–594. doi: 10.1001/archneur.1994.00540180063015

Botelho, C., Teixeira, F., Rolland, T., Abad, A., and Trancoso, I. (2020). Pathological speech detection using x-vector embeddings. *arXiv* 2003.00864.

Bschor, T., Kühl, K.-P., and Reischies, F. M. (2001). Spontaneous speech of patients with dementia of the Alzheimer type and mild cognitive impairment. *Int. Psychogeriatr*. 13, 289–298. doi: 10.1017/S1041610201007682

Chen, J., Zhu, J., and Ye, J. (2019). "An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech," in *Proceedings of Interspeech 2019* (Graz), 4085–4089. doi: 10.21437/Interspeech.2019-2872

Chien, Y.-W., Hong, S.-Y., Cheah, W.-T., Yao, L.-H., Chang, Y.-L., and Fu, L.-C. (2019). An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network. *Sci. Rep*. 9:19597. doi: 10.1038/s41598-019-56020-x

Choi, H. (2009). Performances in a picture description task in Japanese patients with Alzheimer's disease and with mild cognitive impairment. *Commun. Sci. Disord*. 14, 326–337.

Cooper, P. V. (1990). Discourse production and normal aging: performance on oral picture description tasks. *J. Gerontol*. 45, P210–P214. doi: 10.1093/geronj/45.5.P210

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., et al. (2020). "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," in *Proceedings of Interspeech 2020* (Shanghai), 2182–2186. doi: 10.21437/Interspeech.2020-2635

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J. Alzheimers Dis*. 78, 1547–1574. doi: 10.3233/JAD-200888

De Roeck, E. E., De Deyn, P. P., Dierckx, E., and Engelborghs, S. (2019). Brief cognitive screening instruments for early detection of Alzheimer's disease: a systematic review. *Alzheimers Res. Ther*. 11:21. doi: 10.1186/s13195-019-0474-3

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* 1810.04805.

Edwards, E., Dognin, C., Bollepalli, B., and Singh, M. (2020). "Multiscale system for Alzheimer's dementia recognition through spontaneous speech," in *Proceedings of Interspeech 2020* (Shanghai), 2197–2201. doi: 10.21437/Interspeech.2020-2781

Farzana, S., and Parde, N. (2020). "Exploring MMSE score prediction using verbal and non-verbal cues," in *Proceedings of Interspeech 2020* (Shanghai), 2207–2211. doi: 10.21437/Interspeech.2020-3085

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis*. 49, 407–422. doi: 10.3233/JAD-150520

Giles, E., Patterson, K., and Hodges, J. R. (1996). Performance on the boston cookie theft picture description task in patients with early dementia of the Alzheimer's type: missing information. *Aphasiology* 10, 395–408. doi: 10.1080/02687039608248419

Goodglass, H., and Kaplan, E. (1983). *Boston Diagnostic Aphasia Examination Booklet*. Philadelphia, PA: Lea & Febiger.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., and Hoffmann, I. (2019). Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using asr and linguistic features. *Comput. Speech Lang*. 53, 181–197. doi: 10.1016/j.csl.2018.07.007

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). "Learning word vectors for 157 languages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. (Miyazaki).

Guo, Z., Ling, Z., and Li, Y. (2019). Detecting Alzheimer's disease from continuous speech using language models. *J. Alzheimers Dis*. 70, 1163–1174. doi: 10.3233/JAD-190452

Haider, F., De La Fuente, S., and Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE J. Select. Top. Signal Process*. 14, 272–281. doi: 10.1109/JSTSP.2019.2955022

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., and Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimers Dement*. 10, 260–268. doi: 10.1016/j.dadm.2018.02.004

Hong, S.-Y., Yao, L.-H., Cheah, W.-T., Chang, W.-D., Fu, L.-C., and Chang, Y.-L. (2019). "A novel screening system for Alzheimer's disease based on speech transcripts using neural network," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari: IEEE), 2440–2445. doi: 10.1109/SMC.2019.8914628

Khodabakhsh, A., Yesil, F., Guner, E., and Demiroglu, C. (2015). Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. *EURASIP J. Audio Speech Music Process*. 2015:9. doi: 10.1186/s13636-015-0052-y

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's Dement*. 1, 112–124. doi: 10.1016/j.dadm.2014.11.012

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). "Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition," in *Proceedings of Interspeech 2020* (Shanghai), 2217–2221. doi: 10.21437/Interspeech.2020-3153

Land, W. H., and Schaffer, J. D. (2020). "Alzheimer's disease and speech background," in *The Art and Science of Machine Intelligence* (Cham: Springer), 107–135. doi: 10.1007/978-3-030-18496-4_4

Lee, J. L., Burkholder, R., Flinn, G. B., and Coppess, E. R. (2016). *Working With Chat Transcripts in Python*. Technical Report TR-2016–02, Department of Computer Science, University of Chicago.

Liu, L., Zhao, S., Chen, H., and Wang, A. (2020). A new machine learning method for identifying Alzheimer's disease. *Simul. Model. Pract. Theory* 99:102023. doi: 10.1016/j.simpat.2019.102023

López, J. V. E., Tóth, L., Hoffmann, I., Kálmán, J., Pákáski, M., and Gosztolya, G. (2019). "Assessing Alzheimer's disease from speech using the i-vector approach," in *International Conference on Speech and Computer* (Istanbul: Springer), 289–298. doi: 10.1007/978-3-030-26061-3_30

Lueken, U., Seidl, U., Völker, L., Schweiger, E., Kruse, A., and Schröder, J. (2007). Development of a short version of the apathy evaluation scale specifically adapted for demented nursing home residents. *Am. J. Geriatr. Psychiatry* 15, 376–385. doi: 10.1097/JGP.0b013e3180437db3

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). "Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge," in *Proceedings of Interspeech 2020* (Shanghai), 2172–2176. doi: 10.21437/Interspeech.2020-2571

Mackenzie, C., Brady, M., Norrie, J., and Poedjianto, N. (2007). Picture description in neurologically normal adults: concepts and topic coherence. *Aphasiology* 21, 340–354. doi: 10.1080/02687030600911419

MacWhinney, B. (2014). *The CHILDES Project: Tools for Analyzing Talk, Volume I: Transcription Format and Programs*. New York, NY; Hove, ES: Psychology Press.

Martinc, M., and Pollak, S. (2020). "Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia," in *Proceedings of Interspeech 2020* (Shanghai), 2157–2161. doi: 10.21437/Interspeech.2020-2202

Mendez, M. F., and Ashla-Mendez, M. (1991). Differences between multi-infarct dementia and Alzheimer's disease on unstructured neuropsychological tasks. *J. Clin. Exp. Neuropsychol*. 13, 923–932. doi: 10.1080/01688639108405108

Moro-Velazquez, L., Villalba, J., and Dehak, N. (2020). "Using x-vectors to automatically detect Parkinson's disease from speech," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 1155–1159. doi: 10.1109/ICASSP40776.2020.9053770

Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. *J. Clin. Exp. Neuropsychol*. 40, 917–939. doi: 10.1080/13803395.2018.1446513

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv* 1706.08612. doi: 10.21437/Interspeech.2017-950

Nagumo, R., Zhang, Y., Ogawa, Y., Hosokawa, M., Abe, K., Ukeda, T., et al. (2020). Automatic detection of cognitive impairments through acoustic analysis of speech. *Curr. Alzheimer Res*. 17, 60–68. doi: 10.2174/1567205017666200213094513

Ossewaarde, R., Jonkers, R., Jalvingh, F., and Bastiaanse, R. (2019). "Classification of spontaneous speech of individuals with dementia based on automatic prosody analysis using support vector machines (SVM)," in *The Thirty-Second International Flairs Conference* (Sarasota, FL).

Pappagari, R., Cho, J., Moro-Velázquez, L., and Dehak, N. (2020). "Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity," in *Proceedings of Interspeech 2020* (Shanghai), 2177–2181. doi: 10.21437/Interspeech.2020-2587

Pompili, A., Rolland, T., and Abad, A. (2020). "The INESC-ID multi-modal system for the ADReSS 2020 challenge," in *Proceedings of Interspeech 2020* (Shanghai), 2202–2206. doi: 10.21437/Interspeech.2020-2833

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Number CONF* (Big Island, HI: IEEE Signal Processing Society).

Pulido, M. L. B., Hernández, J. B. A., Ballester, M. Á. F., González, C. M. T., Mekyska, J., and Smékal, Z. (2020). Alzheimer's disease and automatic speech analysis: a review. *Expert Syst. Appl*. 150:113213. doi: 10.1016/j.eswa.2020.113213

Qiao, Y., Xie, X.-Y., Lin, G.-Z., Zou, Y., Chen, S.-D., Ren, R.-J., et al. (2020). Computer-assisted speech analysis in mild cognitive impairment and Alzheimer's disease: a pilot study from Shanghai, China. *J. Alzheimers Dis*. 75, 211–221. doi: 10.3233/JAD-191056

Rohanian, M., Hough, J., and Purver, M. (2020). "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," in *Proceedings of Interspeech 2020* (Shanghai), 2187–2191. doi: 10.21437/Interspeech.2020-2721

Sarawgi, U., Zulfikar, W., Soliman, N., and Maes, P. (2020). "Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity," in *Proceedings of Interspeech 2020* (Shanghai), 2212–2216. doi: 10.21437/Interspeech.2020-3137

Searle, T., Ibrahim, Z., and Dobson, R. (2020). "Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech," in*Proceedings of Interspeech 2020* (Shanghai), 2192–2196. doi: 10.21437/Interspeech.2020-2729

Seidl, U., Lueken, U., Thomann, P. A., Kruse, A., and Schröder, J. (2012). Facial expression in Alzheimer's disease: impact of cognitive deficits and neuropsychiatric symptoms. *Am. J. Alzheimers Dis. Other Dement*. 27, 100–106. doi: 10.1177/1533317512440495

Shibata, D., Wakamiya, S., Kinoshita, A., and Aramaki, E. (2016). "Detecting Japanese patients with Alzheimer's disease based on word category frequencies," in *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)* (Osaka), 78–85.

Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). "Deep neural network embeddings for text-independent speaker verification," in *Interspeech* (Stockholm), 999–1003. doi: 10.21437/Interspeech.2017-620

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). "X-vectors: robust DNN embeddings for speaker recognition," in*2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 5329–5333. doi: 10.1109/ICASSP.2018.8461375

Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). "Automated screening for Alzheimer's dementia through spontaneous speech," in *Proceedings of Interspeech 2020* (Shanghai), 2222–2226. doi: 10.21437/Interspeech.2020-3158

Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol*. 29, 24–54. doi: 10.1177/0261927X09351676

Voleti, R., Liss, J. M., and Berisha, V. (2019). A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE J. Select. Top. Signal Process*. 14, 282–298. doi: 10.1109/JSTSP.2019.2952087

Weiner, J., Herff, C., and Schultz, T. (2016). "Speech-based detection of Alzheimer's disease in conversational German," in *Interspeech* (San Francisco, CA), 1938–1942. doi: 10.21437/Interspeech.2016-100

Yadav, V. G. (2019). The hunt for a cure for Alzheimer's disease receives a timely boost. *Sci. Transl. Med*. 11:eaaz0311. doi: 10.1126/scitranslmed.aaz0311

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease," in *Proceedings of Interspeech 2020* (Shanghai), 2162–2166. doi: 10.21437/Interspeech.2020-2516

Conflict of Interest: The 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.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 18 January 2021
doi: 10.3389/fnagi.2020.607449

Using a Discourse Task to Explore Semantic Ability in Persons With Cognitive Impairment

Malin Antonsson[1,2]*, *Kristina Lundholm Fors*[1,2], *Marie Eckerström*[3] *and Dimitrios Kokkinakis*[1,4]

[1] Department of Swedish, Faculty of Arts, University of Gothenburg, Gothenburg, Sweden, [2] Speech and Language Pathology Unit, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, [3] Institute of Neuroscience and Physiology, University of Gothenburg, Sahlgrenska Academy, Gothenburg, Sweden, [4] Centre for Ageing and Health –AgeCap, University of Gothenburg, Gothenburg, Sweden

This paper uses a discourse task to explore aspects of semantic production in persons with various degree of cognitive impairment and healthy controls. The purpose of the study was to test if an in-depth semantic analysis of a cognitive-linguistic challenging discourse task could differentiate persons with a cognitive decline from those with a stable cognitive impairment. Both quantitative measures of semantic ability, using tests of oral lexical retrieval, and qualitative analysis of a narrative were used to detect semantic difficulties. Besides group comparisons a classification experiment was performed to investigate if the discourse features could be used to improve classification of the participants who had a stable cognitive impairment from those who had cognitively declined. In sum, both types of assessment methods captured difficulties between the groups, but tests of oral lexical retrieval most successfully differentiated between the cognitively stable and the cognitively declined group. Discourse features improved classification accuracy and the best combination of features discriminated between participants with a stable cognitive impairment and those who had cognitively declined with an area under the curve (AUC) of 0.93.

Keywords: discourse, mild cognitive impairment, language and aging, machine learning, semantic impairment

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Andrea Marini,
University of Udine, Italy
Aurélie Pistono,
Ghent University, Belgium

****Correspondence:***
Malin Antonsson
malin.antonsson@neuro.gu.se

Received: *17 September 2020*
Accepted: *14 December 2020*
Published: *18 January 2021*

Citation:
Antonsson M, Lundholm Fors K, Eckerström M and Kokkinakis D (2021) Using a Discourse Task to Explore Semantic Ability in Persons With Cognitive Impairment. Front. Aging Neurosci. 12:607449. doi: 10.3389/fnagi.2020.607449

INTRODUCTION

Dementia disorders are neurodegenerative diseases that affect millions of people each year, and the prevalence is still increasing (Scheltens et al., 2016).The most common type of dementia is Alzheimer's disease (AD), and despite extensive ongoing research, little is known about the cause. The development of most dementia disorders is gradual, and cognitive changes are detectable years, and sometimes decades, before dementia is diagnosed (Reisberg and Gauthier, 2008; Ritchie et al., 2015). Subjective cognitive impairment (SCI) and mild cognitive impairment (MCI) are two conditions that have been identified as states preceding dementia (Reisberg and Gauthier, 2008). MCI is characterized as a condition where cognitive decline is observable in at least one cognitive domain, but which does not have a significant interference with a person's daily life (Gauthier et al., 2006). In SCI, which is a common condition in the aging population and is characterized by mild cognitive complaints, no objectively observable cognitive decline is seen (Mendonça et al., 2016). However, previous longitudinal studies report that up to 44% of persons fulfilling the criteria for MCI may return to normal within a year (Gauthier et al., 2006). It is of clinical importance to identify which persons are at risk of cognitive decline and which are likely to remain cognitively stable. If differences in clinical profiles exist between the groups, this could be of help for clinicians diagnosing and planning the care for these groups. At present, there is no gold standard regarding what tasks to use to evaluate language function in persons at risk of developing dementia. However, what is known is that language deficits in general and specifically semantic difficulties are seen early on McCullough et al. (2019), and multiple evaluation methods might be needed to assess changes in language ability (Taler et al., 2020).

In this study we use a discourse task to explore aspects of semantic production in persons with various degrees of cognitive impairment and healthy controls. The purpose is to test if a semantic analysis of a cognitively and linguistically challenging discourse task can be used to differentiate persons with a progressive cognitive decline from those with a stable cognitive impairment. Both quantitative and qualitative measures of semantic ability are used for the purpose of answering this question.

BACKGROUND

Subtle changes in a person's speech or language use may be an early sign of cognitive decline. When a more pronounced cognitive decline, such as dementia, has developed, alterations in syntax, semantics and pragmatics are often present, whereas in milder forms of cognitive decline such as in MCI, predominantly semantic difficulties are seen (see e.g., Taler and Phillips, 2008). Recent studies have also found discourse related features to differentiate between persons with early cognitive impairments and healthy ageing, such as differences in cohesion (Kim et al., 2019) and global coherence (Mazzon et al., 2019). Substantial efforts have been made to identify markers that can be used to predict cognitive decline and that are associated with dementia. Since language data is relatively easy to collect compared to e.g., blood samples and brain imaging, many studies have focused on finding linguistic signs of early cognitive impairment using both qualitative and quantitative measures (for a review see e.g., Mueller et al., 2016) and exploring data both from language tests and continuous speech.

Tests of Semantic Ability

In semantic verbal fluency (SVF) tasks a person is asked to produce as many items as one can from a certain category during 60 s. Although test of verbal fluency tests measure a combination of various cognitive functions and are commonly used to assess both verbal ability and executive control (Shao et al., 2014). SVF are often used for investigating semantic processing and production. Persons with MCI perform worse than healthy controls on SVF tasks, and research suggests that semantic retrieval is impaired (Demetriou and Holtzer, 2017; Linz et al., 2019). A decline in verbal fluency can in fact be seen very early as shown in a series of studies investigating late middle-aged individuals at risk for MCI, where those having "early" MCI had deficits in verbal fluency (Mueller et al., 2015, 2016; Johnson et al., 2018). Furthermore, a decline in semantic fluency in participants at the pre-MCI stage have been seen to predict later progression to MCI and dementia (Loewenstein et al., 2012). Another aspect of semantic ability is confrontation naming, often measured using the Boston Naming Test (BNT; Kaplan and Weintraub, 1983), which consists of 60 images in decreasing order of word frequency. In a recent meta-analysis, Belleville et al. (2017) assessed the predictive accuracy of different cognitive domains and found that in the language domain, confrontational naming (Ahmed et al., 2008; Eckerström et al., 2013) and SVF (Ahmed et al., 2008; Gallagher et al., 2010; Venneri et al., 2011) both yielded high predictive accuracy. Furthermore, numerous studies have shown a relationship between poor baseline performance on semantic word fluency and later development of dementia (Saxton et al., 2004; Auriacombe et al., 2006; Clark et al., 2009). Naming tests are widely used both clinically and in research and have been found to predict the speed of cognitive decline in AD (Carswell, 1999). However, the diagnostic and prognostic utility of these tests may be limited compared to other neuropsychological tests (Taler and Phillips, 2008), and they may not reflect actual ability to communicate and take active part in conversations (Reppermund et al., 2011). Nevertheless, naming tests have been found to correlate with lexical retrieval of nouns in connected speech for persons with aphasia (Herbert et al., 2008).

Quantitative and Qualitative Analyses of Semantic Ability in Discourse

Whereas, quantitative ways of assessing language, such as language tests, have the benefit of being easy to administer and score, analysis of continuous speech, i.e., discourse, is assumed to have a higher sensitivity for detecting subtle linguistic impairments. Analysis of discourse not only allows for a detailed analysis of lexical, semantic, syntactic, and pragmatic features, but also for an analysis of temporal patterns of language production. In previous research, disfluencies (such as pauses, fillers, and false starts) have been studied as a proxy of word finding difficulties, i.e., semantic impairment. In a review, Boschi et al. (2017) conclude that speech in persons with AD is characterized by low speech rate and numerous hesitations. Further, Gayraud et al. (2011) showed that silent pauses, lengthenings, and hesitations are more common in the speech of persons with AD, but there is no increase in filled pauses, which can be interpreted as a lack of signaling speech production difficulties. While pauses may be seen as a symptom of semantic and lexical impairments, Pistono et al. (2019) suggest that pauses may indicate different types of difficulties, as they found that pauses in persons with AD appeared to be predicted by different cognitive functions, depending on the task, and the function of pauses may change as AD progresses (Davis and Maclagan, 2009). In that sense it should be noted that disfluencies are not solely indicative of word finding difficulties: individual differences may be related to verbal intelligence and working memory for example (Engelhardt et al., 2019). Persons with MCI tend to produce longer hesitations (Szatloczki et al., 2015), more pauses (Meilán et al., 2020) and have a lower speech rate (Szatloczki et al., 2015; Meilán et al., 2020). Although it is often concluded that disfluencies are early signs of cognitive decline, Mueller et al.

(2016) found no difference in disfluencies between participants judged as having preclinical (early) MCI and participants who were cognitively healthy. However, in a more recent study by the same group involving more participants they could see that disfluencies in spoken discourse predicted early MCI status and that those with early MCI declined faster in measures of speech fluency than participants who were cognitively stable (Mueller et al., 2018).

Discourse is affected by semantic impairments, and researchers have investigated how aspects of spoken or written discourse are related to cognitive decline. A seminal study in the field, the Nun study (Snowdon et al., 1996), explored narratives in the form of autobiographical essays written by nuns joining a convent. That study, as well as a few other longitudinal studies, have through a prospective or a retrospective analysis linked changes in semantic and lexical content to cognitive decline or development of dementia later in life (Snowdon et al., 1996; Garrard et al., 2004; Farias et al., 2012). However, most studies rely on cross-sectional analysis to explore language features connected to cognitive decline or carry out longitudinal analysis of persons already diagnosed with some type of impairment. A review found that fluency, semantic and speech production outcome measures are most efficient when discriminating persons with MCI from controls (Filiou et al., 2020). These measures were also useful in discriminating MCI and mild AD from controls, whereas syntactic outcome measures were found to be efficient first at mild-moderate stages of the disease, which is consistent with previous studies (Kemper et al., 1993; Ahmed et al., 2013).

Despite the multiple benefits of using a more in-depth qualitative analysis, this is often discarded in a clinical setting due to time constraints. Hence, there is a need for assessment tools for analysis of continuous speech that are easy to use clinically and that can differentiate between persons with cognitive decline and normal ageing. A protocol was developed by Harris et al. (2008), also described in Kiran et al. (2005) and Fleming (2014) to measure the quality of discourse in a task designed to place high demands on executive functioning. They have also developed a protocol for assessing differences in thematic content and used it to differentiate between persons with MCI and controls, with the intent to capture changes in communicative effectiveness. It has been suggested that subtle changes in the overall communicative effectiveness may be early markers of communicative decline, and that the thematic analyses are more efficient and clinically informative than an analysis of linguistic features when evaluating communicative competence (Harris et al., 2008). This type of analysis can be viewed as a pragmatic approach, and includes an analysis of whether the produced information is relevant to the current topic. The inclusion of off-topic information indicates a disruption of discourse, and has been found to have a higher occurrence in discourse of persons with mild AD (Toledo et al., 2018). A higher occurrence was found of a similar type of disruption of coherence, called modalizations, that can be conceptualized as comments or opinions about the speaker's performance during the discourse (Toledo et al., 2018). Whereas the first study using the complex discourse task called the planning task could discriminate between the groups with regards to the thematic analysis (Harris et al., 2008), the more recent study could not (Fleming, 2014). However, both studies could discriminate persons with MCI from persons without cognitive impairment on some type of linguistic analyses, which implies that the task used is complex enough to be used in early stages of cognitive decline.

The purpose of this study is to explore how semantic impairments associated with cognitive deterioration manifest themselves in discourse, and to investigate if measures of semantic content in discourse can be used to distinguish between persons with a stable cognitive impairment (referred to as our cognitively stable group, CS-group), ongoing cognitive decline (referred to as the cognitively declined group, CD-group), and healthy controls (HC-group). To be able to test our methods used to explore semantic production in this type of task, we first needed to know if our groups differ in term of semantic ability. Hence, our first research question concerns this query. Our research questions are:

Does semantic ability (in terms of oral lexical retrieval) as measured on standardized tests differ between persons with cognitive impairment who have cognitively declined, persons with cognitive impairment who are cognitively stable, and a control group?

Do discourse features, in terms of content and disfluencies, differ between persons with cognitive impairment who have cognitively declined, or are cognitively stable in comparisons with a control group?

Can semantically related discourse features be used to improve classification accuracy when combined with SVF results in a machine learning experiment?

Our hypotheses are that:

- semantic ability as measured on standardized tests differ between persons with cognitive impairment who have cognitively declined, persons with cognitive impairment who are cognitively stable, and a control group. We expect the persons with cognitive impairment who have cognitively declined to score lower on the tests than the persons with cognitive impairment who are cognitively stable, and we expect the control group to score the highest.
- discourse features differ between persons with cognitive impairment who have cognitively declined, persons with cognitive impairment who are cognitively stable, and a control group. We expect the persons with cognitive impairment who have cognitively declined to perform worse with regards to discourse features than the persons with cognitive impairment who are cognitively stable, and we expect the control group to perform best.
- classification accuracy can be improved by adding discourse features to SVF results in a machine learning experiment.

METHOD

Participants

The participants in the study consist of 40 persons with cognitive impairment and 28 healthy controls (HC). The participants with cognitive impairment were recruited from the Gothenburg

TABLE 1 | Demographic data.

	CD ($n = 13$) mean (SD)	CS ($n = 27$) mean (SD)	HC ($n = 28$) mean (SD)	Comparison
Age	74.9 (3.6)	68.6 (6.8)	69.5 (7.3)	$p = 0.013$*
Education	14.9 (3.9)	14.6 (3.0)	13.3 (3.4)	$p = 0.085$
MMSE (max 30)	25.8 (3.3)	28.8 (1.9)	29.2 (1.0)	$p \leq 0.001$***
BNT (max 58)	44.3 (11.1)	53.8 (3.1)	52.9 (3.9)	$p = 0.002$**
SVF	17.6 (4.6)	27.0 (6.2)	25.1 (6.2)	$p \leq 0.001$***

**sig. at p-level 0.05, **sig. at p-level 0.01, ***sig. at p-level 0.001. Note: Two persons declined testing with BNT resulting in n 12 in the CD-group and n 26 in the CS-group in this comparison.*

MCI study, a longitudinal study investigating dementia disorders in patients seeking medical care at a memory clinic (Wallin et al., 2016). Inclusion criteria included age 50–79 years and Swedish as their first and only language before the age of 5 years. Exclusion criteria were occurrence of other health conditions that might affect cognitive functioning, such as stroke or brain tumor, substance abuse, serious psychiatric impairment, major depression, or neurological disease. Additional reasons for exclusion were dyslexia and any uncorrected vision or hearing difficulties. The control group was recruited primarily through senior citizens' organizations, using the same exclusion criteria. They also underwent an assessment to rule out any subjective or objective cognitive impairment, and were excluded if they had a Mini Mental State Examination (MMSE; Folstein et al., 1975) score below 26. An overview of the participants is presented in **Table 1**, together with their scores on the MMSE, BNT (Kaplan and Weintraub, 1983), and SVF.

Data Collection

The data collection was divided into two parts: the neuropsychological testing and cognitive/functional assessments, and the language tasks. The cognitive/functional assessment and the neuropsychological testing was administered at the memory clinic by a psychologist or a supervised research nurse. All testing was then assessed by a psychologist (ME). The examination was performed in two sessions of 1.5–2 h. Neuropsychological testing and cognitive assessment was carried out before the collection of language data and again after the language data collection had been completed.

Participants took part in collection of language data at two dates ∼18 months apart, and this study is based on data from the second data collection. The administration of the language tasks took place in a quiet lab environment at University of Gothenburg. The participants completed a discourse task, the SVF as well as some additional tasks not analyzed in the present study.

The first collection of language data included 91 participants, of which 55 persons were diagnosed with some type of cognitive impairment (MCI or SCI) and 36 HC matched for age and education. At the second collection of language data 21 persons failed to return for various reasons. Additionally, one person was excluded due to poor sound quality in the recordings of the language tasks and HC person was excluded due to an MMSE score below 26 at the renewed cognitive assessment.

Neuropsychological Testing and Assessment of Cognitive Status

All participants underwent neuropsychological testing. The participants with cognitive impairment also underwent a cognitive/functional assessment to determine the level of impairment. The tests were selected by clinical neuropsychologists at the memory clinic based on the tests' documented ability to predict subsequent dementia (Eckerström et al., 2013), and with the aim to cover a broad cognitive spectrum. The level of cognitive impairment was assessed with the Global Deterioration Scale (GDS-scale; Auer and Reisberg, 1997) based on four instruments: MMSE (Folstein et al., 1975), Clinical dementia rating (CDR), Stepwise comparative status analysis (Wallin et al., 1996), and I-FLEX (short version of Executive interview EXIT; Royall et al., 1992).

The neuropsychological test battery included tests of learning and memory, language, attention, and executive function. For learning and memory, Rey Auditory Verbal Learning Test (Geffen et al., 1994), Rey Complex Figure (Meyers and Meyers, 1995), recalled after 3 and 20 min, and Weschler Logical Memory subtest (Wechsler, 2003) were used. For language, Boston Naming Test (Kaplan and Weintraub, 1983), verbal fluency for letters F-A-S (Lezak et al., 2012), similarities subtest from the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 2003) and the Token Test, part 5 (De Renzi and Vignolo, 1962) were used. For attention WAIS Digit Span test, WAIS Digit-Symbol test (Wechsler, 2003), the Trail-Making Test forms A and B (Reitan and Wolfson, 1985), for visuo-spatial ability WAIS Block Design test (Wechsler, 2003), Rey Complex Figure copy, and Silhouettes subtest from the Visual Object and Space Perception Battery (Binetti et al., 1996) were used. Finally, for executive function WAIS Letter-Number sequencing subtest, Parallel Serial Mental Operations (Lezak et al., 2012), and the Stroop test (Regard, 1981) were used. All testing was then assessed by a psychologist (ME).

After the second cognitive assessment, the participants with cognitive impairment were divided into those who had deteriorated since the first assessment, the cognitive decline group (CD, n 13) and those who not had deteriorated, the cognitively stable group (CS, n 27). This categorisation was based both on the cognitive assessment and the neuropsychological testing. Six patients converted from mild cognitive impairment to dementia (i.e., scored GDS 3 at baseline and GDS 4/4+ at follow-up). Another seven patients declined cognitively during the study time, based on neuropsychological testing, but did not fulfill criteria for dementia. When analysing neuropsychological test scores, the cut-off for "cognitively impaired" was set at 1.5 standard deviations below the normal mean. Patients had to score below cut-off on at least one out of the nine test variables. The normal mean scores were calculated based on scores from cognitively healthy volunteers included in the Gothenburg MCI study ($n = 117$), and were controlled for significant differences based on age and years of education. Cognitive decline was based on each patient's number of test variables in the normal vs non-normal range (i.e., using the 1.5 standard deviations cut-off). Cognitive decline was defined as a decline (i.e., changed score from normal to below-normal range) from baseline to follow-up in two or more neuropsychological test variables.

Tests of Semantic Ability

The performance on the SVF with the category "animals" (part of the language data collection) and the BNT (Kaplan and Weintraub, 1983) (part of the neuropsychological tests) were used as baseline measures of semantic ability. Administration and scoring was done in accordance with (Tallberg et al., 2008) for SVF and (Tallberg, 2005) for BNT. Due to inconsistent scoring on two items in BNT, these two items were excluded resulting in a maximum of 58 points instead of 60.

Discourse Task

The spontaneous language material analyzed in the present study consists of a spoken discourse task, which was modeled on the "Trip to New York" task developed and validated by Kiran et al. (2005), and described in Harris et al. (2008). For the purposes of this project, the task was changed to "Trip to Stockholm." The participants were asked to describe how they would prepare for and execute a trip to Stockholm. The instructions were as follows:

Now you are going to do a task where you are asked to think and plan aloud. Imagine that you are going on a vacation a week from now. You are traveling to Stockholm for a 2-week stay. Think about all you will have to do to get ready to go, such as how you will get there, what you will bring, and what you will do. I want you to tell me all of your plans until I ask you to stop after about 5 min.

A few follow-up questions were posed if they had not mentioned this information in their narratives, such as: Who will take care of your mail? What will you bring on your trip? The planning task was designed to elicit connected language, that required the participant to supply conceptual and semantic content related to the cognitive-linguistic schema for travel (Harris et al., 2008). It is further suggested to be complex enough to reveal subtle changes in persons with brain damage, due to its demands on executive functions such as initiation, planning, temporal organization and flexibility, and also semantic, episodic and working memory processes.

Data Preparation

The recordings were transcribed orthographically by two certified speech-language pathologists who transcribed approximately half of the recordings each. The transcribers were instructed to segment the discourse into sentences. A clause was defined as having to contain one finite verb, and a sentence defined as consisting of one or several clauses. Besides considering the clauses, the segmentation was based on the speakers' prosodic markers that could indicate sentence breaks. For example, falling intonation could indicate the end of an utterance and thus marked a sentence break. The transcribers trained together before transcribing the participants' recordings to ensure that they interpreted the transcription key correctly. Additionally, each recording was checked twice by one of the transcribers (the first author).

TABLE 2 | Comparison of basic narrative characteristics for the discourse task per group.

	CD mean (SD)	CS mean (SD)	HC mean (SD)	Comparison
N words	334 (122.9)	439 (201.4)	412 (161.7)	$p = 0.187$
N full sentences	22.4 (7.3)	28.7 (12.5)	29.8 (12.0)	$p = 0.1$
N words per sentences	14.0 (4.3)	14.7 (4.1)	14.1 (5.0)	$p = 0.653$
Total phonation duration	107.1 (36.2)	132.6 (61.8)	117.7 (46.7)	$p = 0.512$

*Note: *sig. at p-level 0.05, **sig. at p-level 0.01, ***sig. at p-level 0.001.*

To make the linguistic analysis more efficient, methods from the field of language technology were used. The transcriptions were annotated with part-of-speech (POS) tags and each word was lemmatized using Sparv (Borin et al., 2016). Alignment of the audio recordings and transcriptions was made using Webmaus (Kisler et al., 2017), with post-corrections done manually.

Linguistic Analyses of Discourse Task

The discourse task was analysed with regard to themes and disfluencies, as described in the following sections. Furthermore, some basic narrative characteristics are presented in **Table 2**. Total phonation duration is the total time spent speaking excluding silent pauses.

Semantic Content

To capture semantic aspects of discourse, we focused on thematic content and modalizing language. Modalizations are sometimes referred to as metadiscourse and can be described as remarks on the content of the story e.g., "yeah I can't think of anything else at the moment that I want to do,"[1] and/or concerns about its production (Farias et al., 2012; Toledo et al., 2018) e.g., "...but I always forget what it is called"[1] or "no by the way that's not correct."[1] The thematic coding was based on a previously validated protocol (Harris et al., 2008) used in several studies on the same population (Kiran et al., 2005; Harris et al., 2008; Fleming, 2014). The coding protocol consists of 13 defined core elements i.e., different subtopics/themes: temporal, transportation/ticket, work school/family, money/cost, clothing/packing, lodging, medication/health, securing/housing, activities, food, people, identification, and local cost/money. These were rated 0 if not mentioned, 1 if mentioned briefly, and 2 if elaborated upon. Verbosity or irrelevant comments resulted in a deduction: −1 if minimally present and −2 if significantly present. Minimally present was defined as one irrelevant comment and significantly present was defined as several irrelevant comments or a longer segment of irrelevant information or verbosity. If a theme was mentioned only after the participant was asked a question about that theme, no point was given. Besides scoring the texts according to Fleming (2014), additional analyses of the themes included analysing the number and proportion of words coded as themes, words

[1] Examples taken from the data (translated from Swedish).

coded as modalizations and words coded as unrelated speech (i.e., irrelevant comments).

Disfluencies

Disfluencies are related to the process of planning and producing language. Four types of disfluencies were annotated and analyzed: silent pauses, fillers, false starts, and self-interrupted sentences. Silent pauses were defined as an interval >120 ms within the discourse that is not filled with speech or other sounds produced by the speakers, such as coughing or laughing. The 120 ms cut-off was chosen based on the detection threshold for acoustic silences in speech (Heldner, 2011). Fillers were defined as sounds that indicate e.g., hesitation or planning but that do not have lexical content. Examples of fillers include "uh" and "um." A false start means that the person has started articulating a word, but did not complete it; e.g., the persons says "I pa- pack shoes."[1] Self-interrupted sentences are sentences where the person started producing a sentence but did not complete it; e.g., the person says "and then you could take some—maybe there is some sightseeing- e thing with bus or something like that."[1] If several disfluencies occurred in a row, they were handled as separate instances. The number of disfluencies present in the speech of the participants were measured, as well as the duration of pauses and fillers.

Classification Experiment

To evaluate the usefulness of the extracted features, we tested whether adding them to the SVF score in a machine learning model would improve classification of participants as cognitively stable or cognitively deteriorating. The classification experiment was implemented in Python and Scikit-learn (Pedregosa et al., 2011). For the classification experiment, three common machine learning models used for supervised classification were used: Support Vector Machines (SVM), Gaussian NaiveBayes (NB), and Logistic Regression (LR). Feature selection was performed with SelectKBest, which keeps the n highest scoring features based on an evaluation with an ANOVA. Leave-one-out cross-validation was used for all models. Features were standardized according to the training set in each fold (except for NaiveBayes, since it is invariant to feature scaling), and default hyper-parameters were used. For evaluation, we use area under the receiver operating characteristics curve (AUC). The AUC is calculated by plotting sensitivity (true positive rate) against false positive rate (1 – specificity), as the decision threshold of the classifier is varied. The area under the resulting curve is the AUC, and the better the model is at classifying the groups, the higher is the resulting AUC.

Statistical Methods

Non-parametric tests were chosen as the groups were rather small, and many of the variables were skewed. Kruskal-Wallis were used to compare differences between the groups and Mann-Whitney U-for independent samples were used for *post-hoc* analyses. A more stringent significance level was adopted due to multiple comparisons. After the Bonferroni corrections the new alpha level was $p = 0.01$ for the comparisons of the lexical features (the basic narrative characteristics presented in **Table 2**), $p = 0.006$ for comparisons of thematic content and modalizations and $p = 0.006$ for comparisons of disfluencies. We chose to report both at a significance level of $p = 0.05$ and at the Bonferroni-corrected level. Since there was a significant difference between the groups in age, where the CD group was significantly older than the other two groups, age was added as a covariate in a univariate linear model (ANCOVA) to explore the effect of age. This was only done when there was a relationship between age and the tested variable. Since ANCOVA is a parametric test the dependent variables were logtransformed to meet the assumption of normality. IBM SPSS Statistics version 25 and 26, and R version 3.6.1 (R Core Team, 2019) were used as computational tools.

Ethical Considerations

The present study is covered by the ethical approval (reference number: 206–16, 2016; T021-18) issued by the regional ethical review board in Gothenburg for a larger project. The participants were informed that they could withdraw their participation at any time. All data was coded and made anonymous.

RESULTS

Tests of Semantic Ability

There were significant differences between the groups on both BNT and SVF, see **Table 1**. *Post-hoc* analyses revealed that the CD group had a significantly lower result than the other two groups on both tests. An ANCOVA was performed to explore the effect of age on the results. Both comparisons were still significant after adjusting for age: BNT $F_{(2,62)} = 7.48$, $p = 0.002$, SVF $F_{(2,64)} = 8.21$, $p = 0.001$.

Analysis of Discourse Task

Basic narrative characteristics of the discourse task are provided for groups in **Table 2**. The groups did not differ significantly on the number of words and sentences produced or on total phonation duration.

Semantic Content

The difference between the groups in the thematic content score was borderline significant (see **Table 3** for an overview of all comparisons related to the thematic analysis) A *post-hoc* analysis revealed a difference between the CD group and HC group ($U = 99.5$, $p = 0.019$), but not between the CD group and the CS group ($U = 116$, $p = 0.08$). Since the thematic content score correlated with age, an ANCOVA was performed with the CD group and the controls added as independent variables, to evaluate the effect of age. Age had a significant effect whereas no effect was seen on the group variable [$F_{(1,38)} = 2.20$; $p = 0.15$], suggesting that age and not group explained the difference in thematic content score in this comparison. The number of words in themes were significantly different between the groups (at level $p < 0.05$), but not in the comparison of the proportion of words in themes, indicating that when adjusting for the total number of words in each narrative the proportion of how much they talked about

TABLE 3 | Comparisons of thematic content and modalizations.

	CD mean (SD)	CS mean (SD)	HC mean (SD)	Comparison
Thematic content score	8.2 (2.5)	9.8 (2.8)	10.6 (2.9)	$p = 0.052$
Words in themes (*n*)	192.8 (103.6)	179.3 (34.5)	290.1 (134.8)	$p = 0.046^{*}$
Words in themes (%)	76.4 (23.0)	87.6 (12.8)	89.4 (10.4)	$p = 0.156$
Modalizations (*n*)	0.8 (0.6)	1.0 (0.9)	1.1 (0.9)	$p = 0.642$
Modalizations (*n* of words)	6.6 (8.2)	11.0 (13.2)	11.4 (14.1)	$p = 0.343$
Modalizations (%)	3.1 (4.2)	3.9 (5.6)	4.1 (4.1)	$p = 0.501$
Unrelated speech (*n*)	0.46 (0.66)	0.07 (0.39)	0.18 (0.61)	$p = 0.013^{*}$
Unrelated speech (*n* of words)	30.3 (46.3)	1.6 (8.5)	9.9 (32.9)	$p = 0.009^{*}$
Unrelated speech (%)	13.0 (23.3)	0.4 (2.2)	2.0 (6.7)	$\mathbf{p = 0.006}^{*}$

*Note: *sig. at p-level 0.05, **sig. at p-level 0.01, ***sig. at p-level 0.001. Bold type numbers are significant at the Bonferroni corrected alpha level p ≤ 0.006.*

TABLE 4 | Analysis of disfluency features.

	CD mean (SD)	CS mean (SD)	HC mean (SD)	Comparison
Silent pauses per 100 words	18.8 (7.7)	14.8 (3.0)	12.5 (3.5)	$p = 0.01^{**}$
Fillers per 100 words	2.6 (1.6)	3.9 (2.3)	3.1 (2.0)	$p = 0.203$
False starts per 100 words	0.54 (1.6)	0.70 (0.70)	1.0 (1.1)	$p = 0.220$
Self-interrupted sentences per total sentences	0.18 (0.12)	0.17 (0.10)	0.14 (0.10)	$p = 0.294$
Disfluencies per 100 words	23.1 (8.3)	20.5 (3.9)	17.7 (5.1)	$p = 0.017^{*}$
Mean pause length (>120 ms)	0.78 (0.21)	0.74 (0.18)	0.71 (0.30)	$p = 0.138$
Maximum pause length (>120 ms)	3.5 (1.7)	3.0 (1.5)	2.4 (1.3)	$p = 0.007^{**}$
Mean filler length	0.45 (0.16)	0.50 (0.15)	0.43 (0.1)	$p = 0.151$

*Note: *sig. at p-level 0.05, **sig. at p-level 0.01, ***sig. at p-level 0.001. Bold type numbers are significant at the Bonferroni corrected alpha level p ≤ 0.006.*

the trip planning was similar. None of the features measuring modalizing speech differed between the groups. Unrelated speech was used rarely, less than once per participant (CD $M = 0.46$, CD $M = 0.07$, and HC $M = 0.18$), but most often for the CD group. A *post-hoc* analysis revealed that the CD group produced a higher proportion of unrelated speech than the CS group ($U = 112$, $p = 0.003$), and the HC group ($U = 129$, $p = 0.032$). Only the difference in proportion of unrelated speech survived the Bonferroni corrected p-level.

Disfluencies

Disfluencies in the narratives of the participants were analyzed, and results (see **Table 4**) showed that the groups differ significantly with regard to the number of pauses used (normalized by number of words), the maximum length of pauses and on the total number of disfluencies, i.e., silent pauses, fillers, false starts and self-interruptions, used (normalized by number of words). However, none of the significant results survived a Bonferroni correction.

TABLE 5 | Area under the curve results for the 3 machine learning algorithms with different combinations of features.

	SVM	LR	NB
SVF	0.86	0.82	0.78
SVF + lexical features	0.86	0.89	0.87
SVF + semantic features	0.87	0.89	0.87
SVF + disfluency features	**0.93**	0.91	0.90
SVF + all features	0.87	0.89	0.87

The boldfaced number indicates the best result. SVF, semantic verbal fluency; SVM, Support Vector Machines; LR, Logistic Regression; NB, Gaussian NaiveBayes.

Post-hoc analyses show that persons with cognitive decline and persons who were cognitively impaired but stable did not differ from each other with regard to any of the significant disfluency measures. However, both groups differed significantly from the healthy controls.

Classification Experiment

We evaluated the predictive accuracy of different collections of features by using them in a machine learning model. Since we in this experiment were interested in separating the CS-group from the CD-group) only features from these two groups were applied in the model. The results are presented in **Table 5**. As a baseline, we trained the model using only the results from the SVF, as impairments on the SVF have been found to be predictive of dementia (Taler and Phillips, 2008; Belleville et al., 2017). Using only this feature, we achieved a best result of $AUC = 0.86$ with SVM. We then added the lexical content features, the semantic features and the fluency features separately to the SVF results, and found that this led to improved results, except when training on the SVF and the lexical features and using the SVM classifier, which gave the same AUC as only training on the SVF. Finally, we trained a model using all features combined. The features found most useful were a combination of SVF results and the disfluency features, and training on this data gave the best results for all three classifiers, with the SVM achieving the highest AUC result of 0.93.

DISCUSSION

The present study aimed to investigate semantic aspects of discourse produced by persons who declined cognitively, were cognitively impaired but stable, and healthy controls. To further capture their semantic production, quantitative measures of semantic ability were assessed with tests of oral lexical retrieval. These methods were used in order to explore which measures that best could discriminate between the groups. In sum, both types of assessment methods captured differences between the groups, but the tests of oral lexical retrieval most successfully differentiated between the cognitively stable and the cognitively declined group. This supports previous research which has shown that especially the SVF is a robust predictor of cognitive decline (Taler and Phillips, 2008; Belleville et al., 2017).

To explore semantic aspects of discourse we used a thematic analysis of content (including modalizations and unrelated speech) and an analysis of disfluencies. The elicitation task and the analysis of thematic content were based on the same protocol as Harris et al. (2008) and Fleming (2014). When comparing our CD group with the participants with MCI in Harris et al. (2008) and Fleming (2014), our results are similar to those of Harris et al. (2008) who found that persons with MCI provided less thematic information than the older healthy controls included in the study, and had more irrelevant comments and verbosity. The presence of content not related to the subject or modalizing speech have been found in previous studies investigating discourse in persons with MCI and mild AD (Duong et al., 2003; Drummond et al., 2015; Pistono et al., 2018; Toledo et al., 2018), and is proposed to be related to problems in the semantic-pragmatic component of the language (Drummond et al., 2015). It is further suggested to be a pragmatic ability in AD patients to be able to comment on their communicative production and that it should be viewed as a communicative strength (Duong et al., 2003, Pistono et al., 2018). Why differences in modalizations are not seen in the present study is not clear, but could perhaps be explained by that the use of a more free discourse task did not evoke as many modalizations as picture based task would do. Another possible explanation could be that the present participants' difficulties were too subtle to reveal a difference in modalizing language as seen in previous studies. The proportion of unrelated speech was the only measure that could differentiate between the group who had cognitively declined and the cognitively stable, however, there was a very low occurrence of unrelated speech. The analysis of disfluencies revealed the largest differences between the groups, and we found that the healthy controls tended to use fewer silent pauses, shorter maximum pause lengths and fewer disfluencies in total compared to the cognitively impaired groups. This result is in line with previous research showing that disfluencies are more common in discourse produced by persons with early MCI (Mueller et al., 2018) and in persons with a clinical diagnosis of MCI (Fleming, 2014; Szatloczki et al., 2015; Meilán et al., 2020).

The last research question concerned if the discourse features could improve classification accuracy when combined with SVF results in a machine learning experiment. Our focus here was distinguishing between persons who are cognitively impaired and showing progressive decline, as opposed to persons with stable cognitive impairment. The best classification results were attained by combining the SVF results with the disfluency features, which had a higher AUC (0.93 using Support Vector Machines) than using only SVF. Based on this, we draw the conclusion that the analysis of disfluencies in connected speech provide complementary information to the results on the SVF, possibly because disfluency features do not solely depend on semantic aspects of language but also executive functions which are known to be impaired in MCI (Gauthier et al., 2006).

The task used in the present study was designed to be more cognitively-linguistic challenging, and was added for the second data collections, since previous experiences from using the cookie theft picture as elicitation, suggested that a more challenging task was needed (Lundholm et al., 2018). The planning task was developed by Kiran et al. (2005) with the intent to stimulate connected language instead of more list-like labeling which sometimes can be the case in picture descriptions, and be sensitive to differences in discourse production. Previous studies suggest that task complexity is important when assessing mild impairments as in the case of early AD (Forbes et al., 2002). However, to our knowledge no study has compared the planning task to another type of task, so we can only rely on theoretical assumptions and previous studies concerning the tasks suitability. The thematic analysis was based on the protocol developed by Harris et al. (2008) and consisted of a scoring system where points were given if a certain core element was mentioned. The benefits of this scoring protocol are that it is relatively easy and quick to analyse. A critique might be that it is a bit crude. For that reason, we also chose to analyse how much the participants talked about things related to the themes (or not related to the themes), and how fluently they talked. This seems to complement the scoring, but would be quite cumbersome to implement in the clinic. For at least some of these findings, such as the importance of temporal analysis (disfluencies), they might be implemented in other tasks such as measuring latencies in BNT, or temporally resolved measures on the SVF (Linz et al., 2019). Another adjustment in the present study was the addition of follow-up questions which were posed if certain elements of the trip were not mentioned. Since the information following these questions were prompted and not mentioned spontaneously, we decided to disregard this information in the scoring. This departure from the original protocol means that our results are not completely comparable to previous studies, and we suggest excluding follow-up questions in future studies if the main outcome measure is the score of thematic content.

A drawback of this type of discourse task when used in the clinic is that it requires manual transcription. In some languages it might be possible to use automatic speech recognition, but for Swedish we did not judge the currently available speech recognition alternatives good enough for this purpose. To avoid manual transcription, the test persons can be asked to describe their trip in writing instead, which may be tested in future studies. In the present study, we used methods from language technology and computational linguistics in order to automate some of the analysis and to test if the discourse measures could improve the classification. Studies mixing manual and automated methods seems to be more and more common in this field and can hopefully complement each other (Boschi et al., 2017). Although most studies use manual transcription and segmentation, annotation with part-of-speech taggers and linguistic analyses with for example parsers are often used to make the analysis more efficient and consistent.

A question raised when using this discourse task may be that, if the tests of lexical retrieval were better at discriminating between the groups, then why not use them instead of a discourse task. However, a task that assesses functional language has a higher ecological validity than a psychometric language test (Bastiaanse and Prins, 2004), and can be more challenging, thus more suitable for subtle impairments. Related to that, Drummond et al. (2015) argued that it is often in narrative discourse elderly persons with cognitive deterioration first experience language problems that they perceive as related to impaired memory, such as repetitions or information gaps in their narratives. Furthermore, since tests such as naming and SVF do not address such discourse, deficits occurring in narrative discourse may go undetected. On the contrary, it is also possible that persons with mild difficulties are able to compensate for their problems with lexical retrieval, seen in naming or SVF tests, in a discourse task. However, even if mild word retrieval difficulties do not always lead to anomia, it might lead to an increase in pauses and other types of disfluencies, which was also the case in our data.

One limitation in the present study was the rather small sample, especially in the group with persons who declined cognitively. At the start of the longitudinal project that this study is a part of, participants with either SCI or MCI were included, but due to dropouts the groups ended up rather small at the second point of data collection. The sample size may explain why so few of the comparisons survived Bonferroni adjustments, even though there was a difference in rank seen at alpha level 0.05.

We chose to categorize the participants with a cognitive impairment, according to if they had declined or not from the time when they were included in the project in order to explore which aspects are related to cognitive deterioration. A consequence of this categorisation was that the group with persons who had cognitively declined had a higher age than the persons who were stable and the controls. Since the risk of cognitive impairment increases with age (Unverzagt et al., 2001), it is not surprising that our groups have these demographic characteristics. However, we decided to adjust the comparisons for this factor in those comparisons were there was a relationship between the dependent variable and age. In the case of BNT and SVF, the difference in results were still significant, but not for the difference in thematic content score seen between the CD group and controls.

In sum, the tasks complement each other where the standardized tests provide easy administration and analysis while the planning task offers a more ecologically valid evaluation of spoken language. The tests will indicate which words the persons struggle to find, whereas a discourse task may also reveal what strategies the persons use when experiencing word finding difficulties, and how they are able to compensate. With a larger number of participants, differences between the groups in the discourse task may become more distinct, but differences in communicative efficacy (thematic content score) and fluency seems the most promising variables for future work.

Although the project that this study is a part of is longitudinal, data on the planning task is only available from the second data collection, since it was included later in order to add tasks with a higher complexity. Longitudinal data on this task is needed in order to find out if discourse features such as the ones used in the present study really are useful predictors of cognitive decline.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because they contain sensitive and personally-identifying information. Requests to access the datasets should be directed to Dimitrios Kokkinakis, dimitrios.kokkinakis@svenska.gu.se.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the regional ethical review board in Gothenburg (reference number: 206–16, 2016; T021-18). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KL and DK designed the overall study protocol and collected the data. MA and KL were responsible for developing the research questions for this study, conducting all linguistic analyses, and wrote the first draft of the paper. ME provided the neuropsychological scores and contributed to the clinical interpretation. KL designed and implemented the classification experiments. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research has been funded by Riksbankens Jubileumsfond—The Swedish Foundation for Humanities and Social Sciences, through the grant agreement no: NHS 14-1761:1.

REFERENCES

Ahmed, S., Haigh, A. F., de Jager, C., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain* 136(Pt 12), 3727–3737. doi: 10.1093/brain/awt269

Ahmed, S., Mitchell, J., Arnold, R., Nestor, P. J., and Hodges, J. R. (2008). Predicting rapid clinical progression in amnestic mild cognitive impairment. *Dement. Geriatr. Cogn. Disord.* 25, 170–177. doi: 10.1159/000113014

Auer, S., and Reisberg, B. (1997). The GDS/FAST staging system. *Int. Psychogeriatr.* 9(Suppl. 1), 167–171. doi: 10.1017/S1041610297004869

Auriacombe, S., Lechevallier, N., Amieva, H., Harston, S., Raoux, N., and Dartigues, J. (2006). A longitudinal study of quantitative and qualitative features of category verbal fluency in incident alzheimer's disease subjects: results from the PAQUID study. *Dement. Geriatr. Cogn. Disord.* 21, 260–266. doi: 10.1159/000091407

Bastiaanse, R., and Prins, R. (2004). Review: analysing the spontaneous speech of aphasic speakers. *Aphasiology* 18, 1075–1091. doi: 10.1080/02687030444000534

Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V., and Croteau, J. (2017). Neuropsychological measures that predict progression from mild cognitive impairment to alzheimer's type dementia in older adults: a systematic review and meta-analysis. *Neuropsychol. Rev.* 27, 328–353. doi: 10.1007/s11065-017-9361-5

Binetti, G., Cappa, S. F., Magni, E., Padovani, A., Bianchetti, A., and Trabucchi, M. (1996). Disorders of visual and spatial perception in the early stage of alzheimer's disease. *Ann. N. Y. Acad. Sci.* 777, 221–225. doi: 10.1111/j.1749-6632.1996.tb34422.x

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., and Schumacher, A. (2016). "Sparv: språkbanken's corpus annotation pipeline infrastructure," in *The Sixth Swedish Language Technology Conference (SLTC)* (Umeå: UmeåUniversity), 17–18.

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., and Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: a review. *Front. Psychol*. 8:269. doi: 10.3389/fpsyg.2017.00269

Carswell, L. M. (1999). Prediction of memory and language performance in normal elderly canadians: implications for the assessment of premorbid cognition in early alzheimer's disease. *Dissert. Abstract. Int. B Sci. Eng.* 60:2935. doi: 10.1093/arclin/14.8.670a

Clark, L. J., Gatz, M., Zheng, L., Chen, Y., McCleary, C., and Mack, W. J. (2009). Longitudinal verbal fluency in normal aging, preclinical, and prevalent alzheimer's disease. *Am. J. Alzheimer's Dis. Other Dement.* 24, 461–468. doi: 10.1177/1533317509345154

Davis, B. H., and Maclagan, M. (2009). Examining pauses in alzheimer's discourse. *Am. J. Alzheimer's Dis. Other Dement.* 24, 141–154. doi: 10.1177/1533317508328138

De Renzi, E., and Vignolo, L. A. (1962). Token test: a sensitive test to detect receptive disturbances in aphasics. *Brain* 85, 665–678. doi: 10.1093/brain/85.4.665

Demetriou, E., and Holtzer, R. (2017). Mild cognitive impairments moderate the effect of time on verbal fluency performance. *J. Int. Neuropsychol. Soc.* 23, 44–55. doi: 10.1017/S1355617716000825

Drummond, C., Coutinho, G., Paz Fonseca, R., Assunção, N., Teldeschi, A., de Oliveira-Souza, R., et al. (2015). Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. *Front. Aging Neurosci.* 7:96. doi: 10.3389/fnagi.2015.00096

Duong, A., Tardif, A., and Ska, B. (2003). Discourse about discourse: what is it and how does it progress in alzheimer's disease? *Brain Cogn.* 53, 177–180. doi: 10.1016/S0278-2626(03)00104-0

Eckerström, C., Olsson, E., Bjerke, M., Malmgren, H., Edman, Å., Wallin, A., et al. (2013). A Combination of neuropsychological, neuroimaging, and cerebrospinal fluid markers predicts conversion from mild cognitive impairment to dementia. *J. Alzheimer's Dis.* 36, 421–431. doi: 10.3233/JAD-122440

Engelhardt, P. E., McMullon, M. E. G., and Corley, M. (2019). Individual differences in the production of disfluency: a latent variable analysis of memory ability and verbal intelligence. *Q. J. Exp. Psychol.* 72, 1084–1101. doi: 10.1177/1747021818778752

Farias, S. T., Chand, V., Bonnici, L., Baynes, K., Harvey, D., Mungas, D., et al. (2012). Idea density measured in late life predicts subsequent cognitive trajectories: implications for the measurement of cognitive reserve. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 67, 677–686. doi: 10.1093/geronb/gbr162

Filiou, R., Bier, N., Slegers, A., Houzé, B., Belchior, P., and Brambati, S. M. (2020). Connected speech assessment in the early detection of alzheimer's disease and mild cognitive impairment: a scoping review. *Aphasiology* 34, 723–755. doi: 10.1080/02687038.2019.1608502

Fleming, V. B. (2014). Early detection of cognitive-linguistic change associated with mild cognitive impairment. *Commun. Disord. Q.* 35, 146–157. doi: 10.1177/1525740113520322

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Forbes, K. E., Venneri, A., and Shanks, M. F. (2002). Distinct patterns of spontaneous speech deterioration: an early predictor of alzheimer's disease. *Brain Cogn*. 48, 356–361. doi: 10.1006/brcg.2001.1377

Gallagher, D., Ni Mhaolain, A., Coen, R., Walsh, C., Kilroy, D., Belinski, K., et al. (2010). Detecting prodromal alzheimer's disease in mild cognitive impairment: utility of the CAMCOG and other neuropsychological predictors. *Int. J. Geriatr. Psychiatry* 25, 1280–1287. doi: 10.1002/gps.2480

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2004). The effects of very early alzheimer's disease on the characteristics of writing by a renowned author. *Brain* 128, 250–260. doi: 10.1093/brain/awh341

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., et al. (2006). Mild cognitive impairment. *Lancet* 367, 1262–1270. doi: 10.1016/S0140-6736(06)68542-5

Gayraud, F., Lee, H., and Barkat-Defradas, M. (2011). Syntactic and lexical context of pauses and hesitations in the discourse of alzheimer patients and healthy elderly subjects. *Clin. Linguist. Phon.* 25, 198–209. doi: 10.3109/02699206.2010.521612

Geffen, G. M., Butterworth, P., and Geffen, L. B. (1994). Test-retest reliability of a new form of the auditory verbal learning test (AVLT). *Arch. Clin. Neuropsychol.* 9, 303–316. doi: 10.1093/arclin/9.4.303

Harris, J. L., Kiran, S., Marquardt, T., and Fleming, V. (2008). Communication wellness check-up©: age-related changes in communicative abilities. *Aphasiology* 22, 813–825. doi: 10.1080/02687030701818034

Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *J. Acoust. Soc. Am.* 130, 508–513. doi: 10.1121/1.3598457

Herbert, R., Hickin, J., Howard, D., Osborne, F., and Best, W. (2008). Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? *Aphasiology* 22, 184–203. doi: 10.1080/02687030701262613

Johnson, S. C., Koscik, R. L., Jonaitis, E. M., Clark, L. R., Mueller, K. D., Berman, S. E., et al. (2018). The wisconsin registry for alzheimer's prevention: a review of findings and current directions. *Alzheimer's Dement.* 10, 130–142. doi: 10.1016/j.dadm.2017.11.007

Kaplan, H., and Weintraub, S. (1983). *The Boston Naming Test, 2nd Edn*. Philadelphia, PA: Lea and Febiger.

Kemper, S., LaBarge, E., Ferraro, F. R., Cheung, H., Cheung, H., and Storandt, M. (1993). On the preservation of syntax in alzheimer's disease: evidence from written sentences. *Arch. Neurol.* 50, 81–86. doi: 10.1001/archneur.1993.00540010075021

Kim, B. S., Kim, Y. B., and Kim, H. (2019). Discourse measures to differentiate between mild cognitive impairment and healthy aging. *Front. Aging Neurosci.* 11:221. doi: 10.3389/fnagi.2019.00221

Kiran, S., Harris, J. L., and Marquardt, T. P. (2005). *Communication Wellness Check-Ups: Age-Related Changes in Communication*. San Diego, CA: ASHA National Convention.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347. doi: 10.1016/j.csl.2017.01.005

Lezak, M. D., Howieson, D., and Loring, D. (2012). *Neuropsychological Assessment, 5th Edn*. Oxford: Oxford University Press.

Linz, N., Lundholm Fors, K., Lindsay, H., Eckerström, M., Alexandersson, J., and Kokkinakis, D. (2019). "Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, eds K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, and K. Loveys (Stroudsburg, PA, USA: Association for Computational Linguistics), 103–113.

Loewenstein, D. A., Greig, M. T., Schinka, J. A., Barker, W., Shen, Q., Potter, E., et al. (2012). An investigation of PreMCI: subtypes and longitudinal outcomes. *Alzheimer's Dementia* 8, 172–179. doi: 10.1016/j.jalz.2011.03.002

Lundholm, F. K., Fraser, K., and Kokkinakis, D. (2018). Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment. *Stud Health Technol Inform*. 247, 705–709 doi: 10.3233/978-1-61499-852-5-705

Mazzon, G., Ajčević, M., Cattaruzza, T., Menichelli, A., Guerriero, M., Capitanio, S., et al. (2019). Connected speech deficit as an early hallmark of CSF-defined alzheimer's disease and correlation with cerebral hypoperfusion pattern. *Curr. Alzheimer Res.* 16, 483–494. doi: 10.2174/1567205016666190506141733

McCullough, K. C., Bayles, K. A., and Bouldin, E. D. (2019). Language performance of individuals at risk for mild cognitive impairment. *J. Speech Lang. Hear. Res.* 62, 706–722. doi: 10.1044/2018_JSLHR-L-18-0232

Meilán, J. J. G., Martínez-Sánchez, F., Martínez-Nicolás, I., Llorente, T. E., and Carro, J. (2020). Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. *Behav. Neurol.* 2020:4683573. doi: 10.1155/2020/4683573

Mendonça, M. D., Alves, L., and Bugalho, P. (2016). From subjective cognitive complaints to dementia. *Am. J. Alzheimer's Dis. Other Dement.* 31, 105–114. doi: 10.1177/1533317515592331

Meyers, J., and Meyers, K. (1995). *Rey Complex Figure Test and Recognition Trial.* San Antonio, TX: The Psychological Corporation.

Mueller, K. D., Koscik, R. L., Hermann, B. P., Johnson, S. C., and Turkstra, L. S. (2018). Declines in connected language are associated with very early mild cognitive impairment: results from the wisconsin registry for alzheimer's prevention. *Front. Aging Neurosci.* 9:437. doi: 10.3389/fnagi.2017.00437

Mueller, K. D., Koscik, R. L., LaRue, A., Clark, L. R., Hermann, B., Johnson, S. C., et al. (2015). Verbal fluency and early memory decline: results from the wisconsin registry for alzheimer's prevention. *Arch. Clin. Neuropsychol.* 30, 448–457. doi: 10.1093/arclin/acv030

Mueller, K. D., Koscik, R. L., Turkstra, L. S., Riedeman, S. K., Larue, A., Clark, L. R., et al. (2016). Connected language in late middle-aged adults at risk for alzheimer's disease HHS public access. *J. Alzheimers. Dis.* 54, 1539–1550. doi: 10.3233/JAD-160252

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in {P}ython. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pistono, A., Jucla, M., Bézy, C., Lemesle, B., Le Men, J., and Pariente, J. (2018). Discourse macrolinguistic impairment as a marker of linguistic and extralinguistic functions decline in early alzheimer's disease. *Int. J. Lang. Commun. Disord.* 54, 390–400. doi: 10.1111/1460-6984.12444

Pistono, A., Pariente, J., Bézy, C., Lemesle, B., Le Men, J., and Jucla, M. (2019). What happens when nothing happens? *An investigation of pauses as a compensatory mechanism in early alzheimer's disease. Neuropsychologia* 124, 133–143. doi: 10.1016/j.neuropsychologia.2018.12.018

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Vienna. Available online at: https://www.r-project.org/

Regard, M. (1981). *Cognitive Rigidity and Flexibility: A Neuropsychological Study*. Victoria: University of Victoria.

Reisberg, B., and Gauthier, S. (2008). Current evidence for subjective cognitive impairment (SCI) as the pre-mild cognitive impairment (MCI) stage of subsequently manifest alzheimer's disease. *Int. Psychogeriatr.* 20, 1–16. doi: 10.1017/S1041610207006412

Reitan, R. M., and Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Therapy and Clinical Interpretation*. Tucson, AZ: Neuropsychological Press.

Reppermund, S., Sachdev, P. S., Crawford, J., Kochan, N. A., Slavin, M. J., Kang, K., et al. (2011). The relationship of neuropsychological function to instrumental activities of daily living in mild cognitive impairment. *Int. J. Geriatr. Psychiatry* 26, 843–852. doi: 10.1002/gps.2612

Ritchie, K., Ritchie, C. W., Yaffe, K., Skoog, I., and Scarmeas, N. (2015). Is late-onset alzheimer's disease really a disease of midlife? *Alzheimer's Dement*. 1, 122–130. doi: 10.1016/j.trci.2015.06.004

Royall, D. R., Mahurin, R. K., and Gray, K. F. (1992). Bedside assessment of executive cognitive impairment: the executive interview. *J. Am. Geriatr. Soc.* 40, 1221–1226. doi: 10.1111/j.1532-5415.1992.tb03646.x

Saxton, J., Lopez, O. L., Ratcliff, G., Dulberg, C., Fried, L. P., Carlson, M. C., et al. (2004). Preclinical alzheimer disease: neuropsychological test performance 1.5 to 8 years prior to onset. *Neurology* 63, 2341–2347. doi: 10.1212/01.WNL.0000147470.58328.50

Scheltens, P., Blennow, K., Breteler, M., de Strooper, B., Frisoni, G., Salloway, S., et al. (2016). Alzheimer's disease. *Lancet* 388, 505–517. doi: 10.1016/S0140-6736(15)01124-1

Shao, Z., Janse, E., Visser, K., and Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front. Psychol.* 5:772. doi: 10.3389/fpsyg.2014.00772

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and alzheimer's disease in late life: findings from the nun study. *JAMA* 275, 528–532. doi: 10.1001/jama.275.7.528

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in alzheimer's disease, is that an early sign? Importance of changes in language abilities in alzheimer's disease. *Front. Aging Neurosci.* 7:195. doi: 10.3389/fnagi.2015.00195

Taler, V., Monetta, L., Sheppard, C., and Ohman, A. (2020). Semantic function in mild cognitive impairment. *Front. Psychol.* 10:3041. doi: 10.3389/fpsyg.2019.03041

Taler, V., and Phillips, N. A. (2008). Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *J. Clin. Exp. Neuropsychol.* 30, 501–556. doi: 10.1080/13803390701550128

Tallberg, I. (2005). The boston naming test in Swedish: normative data. *Brain Lang.* 94, 19–31. doi: 10.1016/j.bandl.2004.11.004

Tallberg, I., Ivachova, E., Jones Tinghag, K., and Östberg, P. (2008). Swedish norms for word fluency tests: FAS, animals and verbs. *Scand. J. Psychol.* 49, 479–485. doi: 10.1111/j.1467-9450.2008.00653.x

Toledo, C. M., Aluísio, S. M., dos Santos, L. B., Dozzi Brucki, S. M., Sturzeneker Trés, E., de Oliveira, M. O., et al. (2018). Analysis of macrolinguistic aspects of narratives from individuals with alzheimer's disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer's Dement* 10, 31–40. doi: 10.1016/j.dadm.2017.08.005

Unverzagt, F. W., Gao, S., Baiyewu, O., Ogunniyi, A. O., Gureje, O., Perkins, A., et al. (2001). Prevalence of cognitive impairment: data from the indianapolis study of health and aging. *Neurology* 57, 1655–1662. doi: 10.1212/WNL.57.9.1655

Venneri, A., Gorgoglione, G., Toraci, C., Nocetti, L., Panzetti, P., and Nichelli, P. (2011). Combining neuropsychological and structural neuroimaging indicators of conversion to alzheimer's disease in amnestic mild cognitive impairment. *Curr. Alzheimer Res.* 8, 789–797. doi: 10.2174/156720511797633160

Wallin, A., Edman, Å., Blennow, K., Gottfries, C., Karlsson, I., Regland, B., et al. (1996). Stepwise comparative status analysis (STEP): a tool for identification of regional brain syndromes in dementia. *J. Geriatr. Psychiatry Neurol.* 9, 185–199. doi: 10.1177/089198879600900406

Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., et al. (2016). The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *J. Cereb. Blood. Flow. Metab*. 36, 114–131. doi: 10.1038/jcbfm.2015.147

Wechsler, D. (2003). *WAIS-III Manual (Swedish Version)*. New York, NY: Harcourt Assessment, Inc.

Conflict of Interest: The 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.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 29 January 2021
doi: 10.3389/fcomp.2020.624488

Pauses for Detection of Alzheimer's Disease

Jiahong Yuan[1]*, *Xingyu Cai*[1], *Yuchen Bian*[1], *Zheng Ye*[2] *and Kenneth Church*[1]

[1]*Baidu Research, Sunnyvale, CA, United States,* [2]*Institute of Neuroscience, Key Laboratory of Primate Neurobiology, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China*

Pauses, disfluencies and language problems in Alzheimer's disease can be naturally modeled by fine-tuning Transformer-based pre-trained language models such as BERT and ERNIE. Using this method with pause-encoded transcripts, we achieved 89.6% accuracy on the test set of the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge. The best accuracy was obtained with ERNIE, plus an encoding of pauses. Robustness is a challenge for large models and small training sets. Ensemble over many runs of BERT/ERNIE fine-tuning reduced variance and improved accuracy. We found that *um* was used much less frequently in Alzheimer's speech, compared to *uh*. We discussed this interesting finding from linguistic and cognitive perspectives.

Keywords: Alzheiemer's disease, pause, BERT, ERNIE, ensemble

OPEN ACCESS

Edited by:
Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:
Jiri Pribil,
Slovak Academy of Sciences, Slovakia
Shane Sheehan,
University of Edinburgh,
United Kingdom

****Correspondence:***
Jiahong Yuan
jiahongyuan@baidu.com

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *31 October 2020*
Accepted: *11 December 2020*
Published: *29 January 2021*

Citation:
Yuan J, Cai X, Bian Y, Ye Z and
Church K (2021) Pauses for Detection
of Alzheimer's Disease.
Front. Comput. Sci. 2:624488.
doi: 10.3389/fcomp.2020.624488

1 INTRODUCTION

Alzheimer's disease (AD) involves a progressive degeneration of brain cells that is irreversible (Mattson, 2004). One of the first signs of the disease is deterioration in language and speech production (Mueller et al., 2017). It is desirable to use language and speech for AD detection (Laske et al., 2015). In this paper, we investigate the use of pauses in speech (both unfilled and filled pauses such as "uh" and "um") for this task.

1.1 Pauses

Unfilled pauses play an important role in speech. The occurrence of pauses is subject to physiological, linguistic, and cognitive constraints (Goldman-Eisler, 1961; Rochester, 1973; Butcher, 1981; Zellner, 1994; Clark, 2006; Ramanarayanan et al., 2013; Hawthorne and Gerken, 2014). How different constraints interact in pause production has been an active research subject for decades. In normal speech, the likelihood of pause occurrence and the duration of pauses are correlated with syntactic and prosodic structure (Brown and Miron, 1971; Grosjean et al., 1971; Krivokapic, 2007). For example, if a sentence has a syntactically complex subject and a syntactically complex object, speakers tend to pause at the subject-verb phrase boundary, and pause duration increases with upcoming complexity (Ferreira, 1991). It has been demonstrated that pauses in speech are used by listeners in sentence parsing (Schepman and Rodway, 2000), and the pause information can benefit automatic parsing (Tran et al., 2018).

Atypical pausing is characteristic of disordered speech such as in Alzheimer's disease, and pauses are often used to measure language and speech problems (Ramig et al., 1995; Yuan et al., 2016; Shea and Leonard, 2019). The difference between typical and atypical pauses is not only on their frequency and duration, but also on where they occur. In this study, we propose a method to encode pauses in transcripts in order to capture the associations between pauses and words through fine-tuning pre-trained language models such as BERT [19] and ERNIE [20], which we describe in **Section 1.2**.

The use of filled pauses may also be different between AD and normal speech. English has two common filled pauses, *uh* and *um*. There is a debate in the literature as to whether *uh* and *um* are intentionally produced by speakers (Clark and Fox Tree, 2002; Corley and Stewart, 2008). From sociolinguistic point of view, women and younger people tend to use more *um* vs. *uh* than men and older people (Tottie, 2011; Wieling et al., 2016). It has also been reported that autistic children use *um* less frequently than normal children (Gorman et al., 2016; Irvine et al., 2016), and that *um* occurs less frequently and is shorter during lying compared to truth-telling (Arciuli et al., 2010).

1.2 Pre-trained LMs and Self-Attention

Modern pre-trained language models such as BERT (Devlin et al., 2018) and ERNIE (Sun et al., 2019) were trained on extremely large corpora. These models appear to capture a wide range of linguistic facts including lexical knowledge, phonology, syntax, semantics and pragmatics. Recent literature is reporting considerable success on a variety of benchmark tasks with BERT and BERT-like models.[1] We expect that the language characteristics of AD can also be captured by the pre-trained language models when fine-tuned to the task of AD classification.

BERT and BERT-like models are based on the Transformer architecture (Vaswani et al., 2017). These models use self-attention to capture associations among words. Each attention head operates on the elements in a sequence (e.g., words in the transcript for a subject), and computes a new sequence of the weighed sum of (transformed) input elements. There are various versions of BERT and ERNIE. There is a base model with 12 layers and 12 attention heads for each layer, as well as a larger model with 24 layers and 16 attention heads for each layer. Conceptually the self-attention mechanism can naturally model many language problems in AD, including repetitions of words and phrases, use of particular words (and classes of words), as well as pauses. By inserting pauses in word transcripts, we enable BERT-like models to learn the language problems involving pauses.

Previous studies have found that when fine-tuning BERT for downstream tasks with a small data set, the model has a high variance in performance. Even with the same hyperparameter values, distinct random seeds can lead to substantially different results. Dodge et al. (2020) conducted a large-scale study on this issue. They fine-tuned BERT hundreds of times while varying only the random seeds, and found that the best-found model significantly outperformed previous reported results using the same model. In this situation, using just one final model for prediction is risky given the variance in performance during training. We propose an ensemble method to address this concern.

1.3 Automatic Detection of AD

There is a considerable literature on AD detection from continuous speech (Filiou et al., 2019; Pulido et al., 2020). This literature considers a wide variety of features and machine learning techniques. Fraser et al. (2016) used 370 acoustic and linguistic features to train logistic regression models for classifying AD and normal speech. Gosztolya et al. (2019) found that acoustic and linguistic features were about equally effective for AD classification, but the combination of the two performed better than either by itself. Neural network models such as Convolutional Neural Networks and Long Short-Term Memory (LSTM) have also been employed for the task (de Ipiña et al., 2017; Fritsch et al., 2019; Palo and Parde, 2019), and very promising results have been reported. However, it is difficult to compare these different approaches, because of the lack of standardized training and test data sets. The ADReSS challenge of INTERSPEECH 2020 is "to define a shared task through which different approaches to AD detection, based on spontaneous speech, could be compared" (Luz et al., 2020). This paper stems from our effort for the shared task.

2 DATA AND ANALYSIS

2.1 Data

The data consists of speech recordings and transcripts of descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Exam (Goodglass et al., 2001). Transcripts were annotated using the CHAT coding system (MacWhinney, 2000). We only used word transcripts, the morphological and syntactic annotations in the transcripts were not used in our experiments.

The training set contains 108 speakers, and the test set contains 48 speakers. In each data set, half of the speakers are people with AD and half are non-AD (healthy control subjects). Both data sets were provided by the challenge. The organizers also provided speech segments extracted from the recordings using a simple voice detection algorithm, but no transcripts were available for the speech segments. We didn't use these speech segments. Our experiments were based on the entire recordings and transcripts.

2.2 Processing Transcripts and Forced Alignment

The transcripts in the data sets were annotated in the CHAT format, which can be conveniently created and analyzed using CLAN (MacWhinney, 2000). For example: "the [x 3] bench [: stool]". In this example, [x 3] indicates that the word "the" was repeated three times [: stool] indicates that the preceding word, "bench" (which was actually produced), refers to stool. Details of the transcription format can be found in (MacWhinney, 2000).

For the purpose of forced alignment and fine tuning, we converted the transcripts into words and tokens that represent what were actually produced in speech. "w [x n]" were replaced by repetitions of w for n times, punctuation marks and various comments annotated between "[]" were removed. Symbols such as (.), (..), (. . .), <, >, / and xxx were also removed.

The processed transcripts were forced aligned with speech recordings using the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008). The aligner used a special model "sp" to

[1] https://gluebenchmark.com

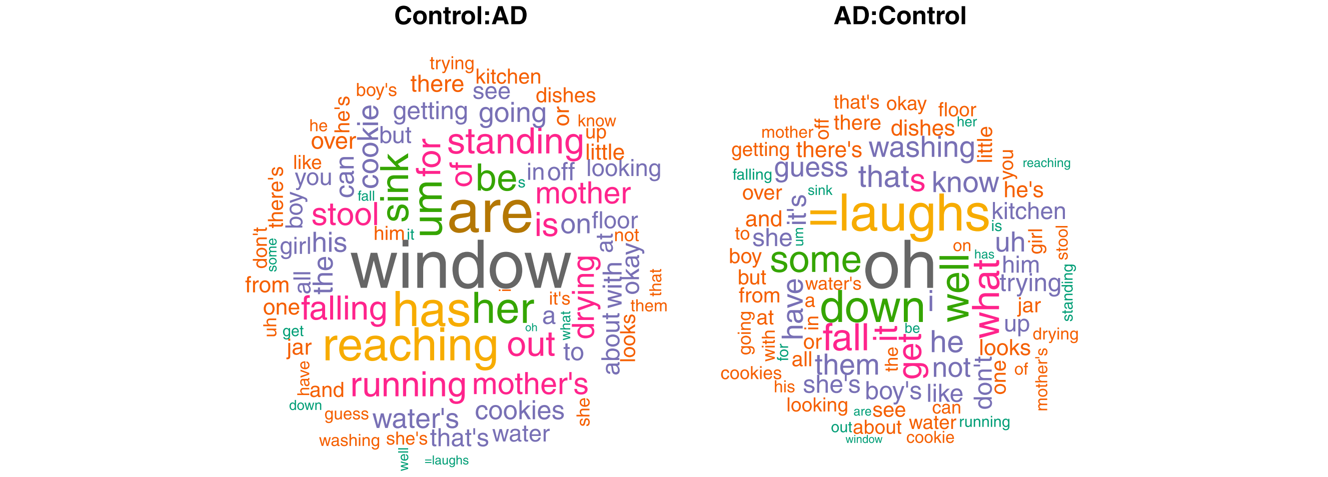


FIGURE 1 | The word cloud on the left highlights words that are more common among control subjects than AD; the word cloud on the right highlights words that are more common among AD than control.

TABLE 1 | Subjects with AD say uh more often, and um less often.

	uh	*um*
Control (non-AD)	130	51
Dementia (AD)	183	20

identify between-word pauses. After forced alignment, the speech segments that belong to the interviewer were excluded. The pauses at the beginning and the end of the recordings were also excluded. Only the subjects' speech, including pauses in turn-taking between the interviewer and the subject, were used.

2.3 Word Frequency and *Uh*/*Um*

From the training data set, we calculated word frequencies for the Control and AD groups respectively. Words that appear 10 or more times in both groups are shown in the word clouds in **Figure 1**. The following words are at least two times more frequent in AD than in Control: *oh* (4.33), = *laughs* (laughter, 3.18), *down* (2.66), *well* (2.42), *some* (2.2), *what* (2.16), *fall* (2.15). And the words that are at least two times more frequent in Control than in AD are: *window* (4.4), *are* (3.83), *has* (3.0), *reaching* (2.8), *her* (2.62), *um* (2.55), *sink* (2.3), *be* (2.21), *standing* (2.06).

Compared to controls, subjects with AD used relatively more laughter and semantically "empty" words such as *oh*, *well*, and *some*, and fewer present particles (*-ing* verbs). This is consistent with findings in the literature. **Table 1** shows an interesting difference for filled pauses. The subjects with AD used more *uh* than the control subjects, but their use of *um* was much less frequent.

2.4 Unfilled Pauses

Duration of pauses was calculated from forced alignment. Pauses under 50 ms were excluded, as well as pauses in the interviewer's speech. We binned the remaining pauses by duration as shown in **Figure 2**. Subjects with AD have more pauses in every group, but the difference between subjects with AD and non-AD is particularly noticeable for longer pauses.

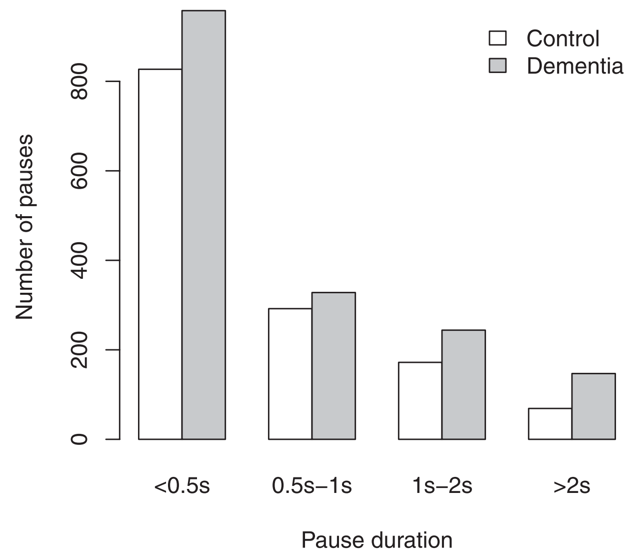
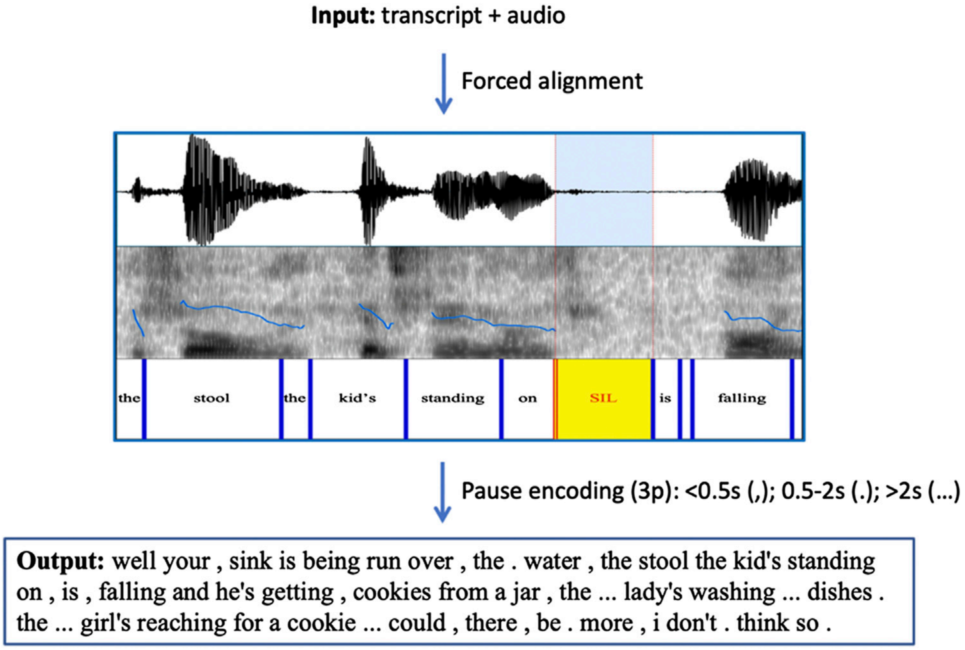


FIGURE 2 | Subjects with AD have more pauses (in all duration bins).

3 BERT AND ERNIE FINE-TUNING

3.1 Input and Hyperparameters

Pre-trained BERT and ERNIE models were fine-turned for the AD classification task. Each of the $N = 108$ training speakers is considered a data point. The input to the


Input: transcript + audio
Forced alignment
the
stool
the
kid's
standing
on
SIL
is
falling
Pause encoding (3p): <0.5s (,); 0.5-2s (.); >2s (...)

Output: well your , sink is being run over , the . water , the stool the kid's standing on , is , falling and he's getting , cookies from a jar , the ... lady's washing ... dishes . the ... girl's reaching for a cookie ... could , there , be . more , i don't . think so .

FIGURE 3 | Procedure for pause encoding.

model consists of a sequence of words from the processed transcript for every speaker (as described in **Section 2.2**). The output is the class of the speaker, 0 for Control and one for AD.

We also encoded pauses in the input word sequence. We grouped pauses into three bins: short (under 0.5 s); medium (0.5–2 s); and long (over 2 s). The three bins of pauses are coded using three punctuations ",", ".", and "...", respectively. Because all punctuations were removed from the processed transcripts, these inserted punctuations only represent pauses. The procedure is illustrated in **Figure 3**.

We used Bert-for-Sequence-Classification[2] for fine-tuning. We tried both "bert-base-uncased" and "bert-large-uncased", and found slightly better performance with the larger model. The following hyperparameters (slightly tuned) were chosen: learning rate = 2e-5, batch size = 4, epochs = 8, max input length of 256 (sufficient to cover most cases). The standard default tokenizer was used (with an instruction not to split "..."). Two special tokens, [CLS] and [SEP], were added to the beginning and the end of each input.

ERNIE fine-tuning started with the "ERNIE-large" pre-trained model (24 layers with 16 attention heads per layer). We used the default tokenizer, and the following hyperparameters: learning rate = 2e-5, batch size = 8, epochs = 20 and max input length of 256.

The fine-tuning process is illustrated in **Figure 4**.

3.2 Ensemble Reduces Variance in LOO Accuracy

When conducting LOO (leave-one-out) cross-validation on the training set, large differences in accuracy across runs were observed. We computed 50 runs of LOO cross-validation. The hyperparameter setting was the same across runs except for random seeds. The results are shown in the last row ($N = 1$) of **Tables 2** and **3**. Over the 50 runs, LOO accuracy ranged from 0.75 to 0.86 for BERT with three pauses, from 0.78 to 0.87 for ERNIE with three pauses, and from 0.77 to 0.85 for ERNIE with no Pauses. The large variance suggests performance on unseen data is likely to be brittle. Such brittleness is to be expected given the large size of the BERT and ERNIE models and the small size of the training set (108 subjects).

To address this brittleness, we introduced the following ensemble procedure. From the results of LOO cross-validation, we calculated the majority vote over N runs for each of the 108 subjects, and used the majority vote to return a single label for each subject. To make sure that the ensemble estimates would generalize to unseen data, we tested the method by selecting $N = 5$, $N = 15$, ..., runs from the 50 runs of LOO cross-validation. The results are shown in **Table 2** and **3**. In the tables, the first row summarizes 100 draws of $N = 5$ runs. The second row is similar, except $N = 15$. All of the ensemble rows have better means and less variance than the last row, which summarizes the 50 individual runs of LOO cross-validation without ensemble ($N = 1$). **Figure 5** illustrates **Table 2** and **3**. In **Figure 5** the black lines represent accuracy of individual runs whereas the purple lines represent ensemble accuracy of $N = 35$. We can see that there is a wide variance in individual runs (black).

[2] https://github.com/huggingface/transformers

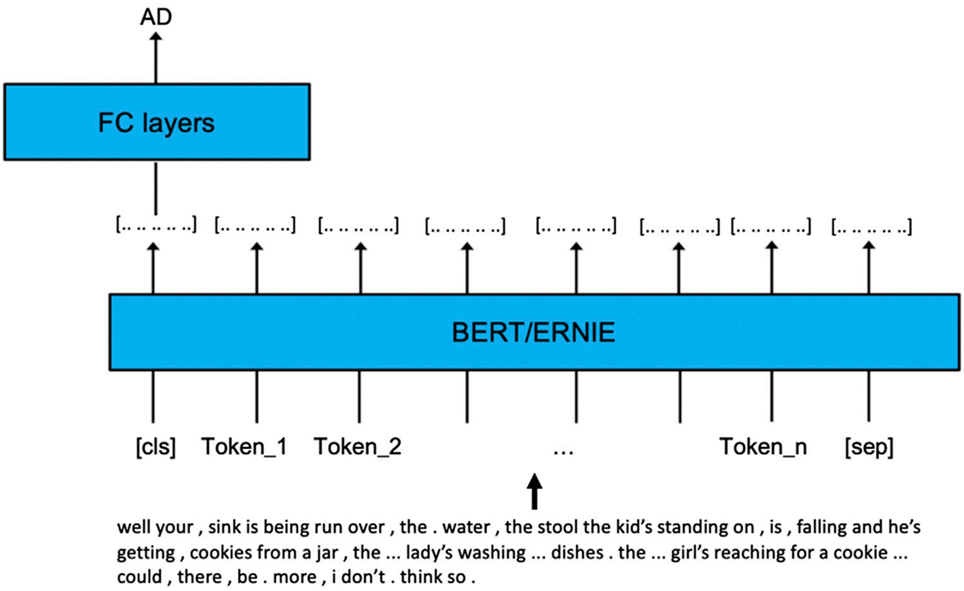

AD
FC layers
[.. .. ..] [.. .. ..] [.. .. ..] [.. .. ..] [.. .. ..] [.. .. ..] [.. .. ..] [.. .. ..]
BERT/ERNIE
[cls] Token_1 Token_2 … Token_n [sep]
well your , sink is being run over , the . water , the stool the kid's standing on , is , falling and he's
getting , cookies from a jar , the ... lady's washing ... dishes . the ... girl's reaching for a cookie ...
could , there , be . more , i don't . think so .

FIGURE 4 | Procedure for fine-tuning.

TABLE 2 | Ensemble improves LOO (leave-one-out) estimates of accuracy; better means with less variance.

	BERT with three pauses	
N	Mean ± sd	min - max
5	0.837 ± 0.010	0.815–0.861
15	0.840 ± 0.011	0.815–0.861
25	0.839 ± 0.011	0.815–0.870
35	**0.838 ± 0.010**	**0.824–0.861**
45	0.839 ± 0.011	0.824–0.861
1	**0.819 ± 0.023**	**0.750–0.861**

TABLE 3 | Ensemble also improves LOO for ERNIE (with and without pauses). LOO results are better with pauses than without, and better with ERNIE than BERT.

	ERNIE with three pauses		ERNIE with No pauses	
N	Mean ± std	Min - max	Mean ± std	Min - max
5	0.845 ± 0.013	0.806–0.880	0.828 ± 0.016	0.796–0.870
15	0.851 ± 0.008	0.833–0.870	0.831 ± 0.012	0.796–0.861
25	0.853 ± 0.007	0.833–0.870	0.833 ± 0.010	0.815–0.861
35	**0.854 ± 0.007**	**0.824–0.861**	**0.836 ± 0.009**	**0.815–0.852**
45	0.854 ± 0.007	0.833–0.861	0.834 ± 0.008	0.815–0.861
1	**0.827 ± 0.020**	**0.778–0.870**	**0.816 ± 0.023**	**0.769–0.852**

The proposed ensemble method (purple) improves the mean and reduces variance over estimates based on a single run.

4 EVALUATION

Under the rules of the challenge, each team is allowed to submit results of five attempts for evaluation. Predictions on the test set from the following five models were submitted for evaluation: BERT0p, BERT3p, BERT6p, ERNIE0p, and ERNIE3p. 0p indicates that no pause was encoded, and 3p and 6p indicate, respectively, that three and six lengths of pauses were encoded. To compare with three pauses, 6p represents six bins of pauses, encoded as: "," (under 0.5 s), "." (0.5–1 s); ".." (1–2 s), ". . ." (2–3 s), ". . . ." (3–4 s), ". . . . ." (over than 4 s). The dots are separated from each other, as different tokens.

Following the method proposed in **Section 3.2**, we made 35 runs of training for each of the five models, with 35 random seeds. The classification of each sample in the test set was based on the majority vote of 35 predictions. **Table 4** lists the evaluation scores received from the organizers.

The best accuracy was 89.6%, obtained with ERNIE and three pauses. It is a nearly 15% increase from the baseline of 75.0% (Luz et al., 2020).

ERNIE outperformed BERT by 4% on input of both three pauses and no pause. Encoding pauses improved the accuracy for both BERT and ERNIE. There was no difference between three pauses and six pauses in terms of improvement in accuracy.

5 DISCUSSION

The group with AD used more *uh* but less *um* than the control group. In speech production, disfluencies such as hesitations and speech errors are correlated with cognitive functions such as cognitive load, arousal, and working memory (Daneman, 1991; Arciuli et al., 2010). Hesitations and disfluencies increase with increased cognitive load and arousal as well as impaired working memory. This may explain why the group with AD used more *uh*, as a filled pause and hesitation marker. More interestingly, they

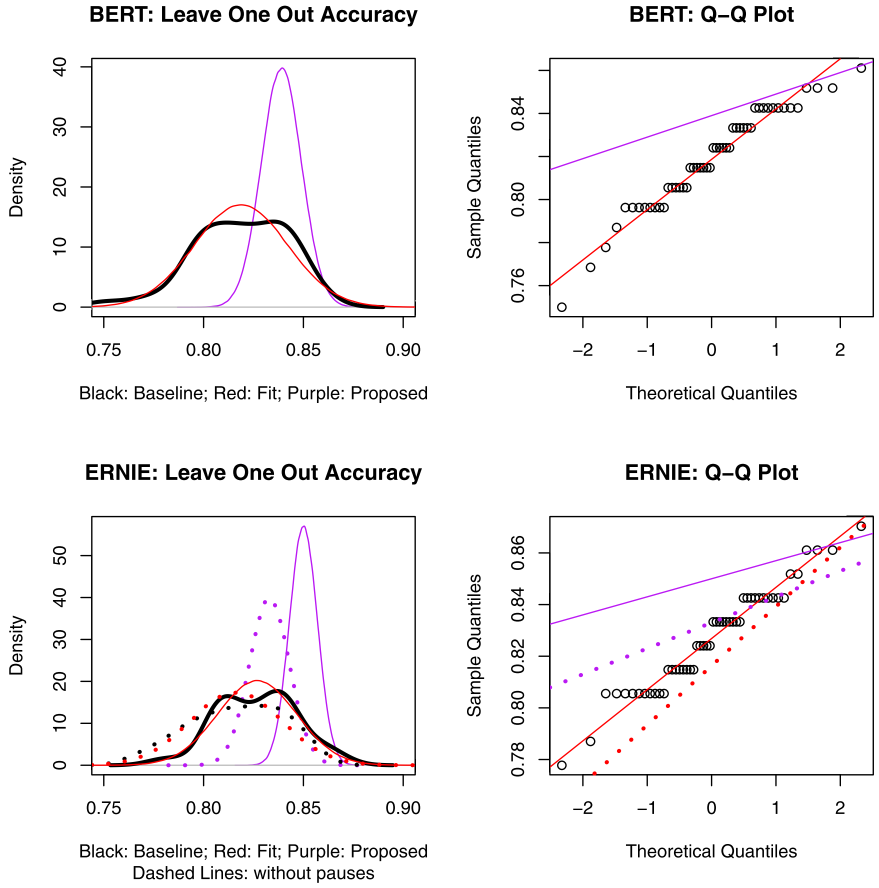


FIGURE 5 | Individual and ensemble Leave-one-out (LOO) accuracy for BERT with pauses (top) and ERNIE with and without pauses (bottom). Black lines represent accuracy of individual runs; purple lines represent ensemble accuracy of $N = 35$.

TABLE 4 | Evaluation results: Best accuracy (acc) with ERNIE and three pauses (3p). Pauses are helpful: three pauses (3p) and six pauses (6p) have better accuracy than no pauses (0p).

	Precision		**Recall**		**F1**		**Acc**
	Non-AD	AD	Non-AD	AD	Non-AD	AD	
Baseline ()	0.700	0.830	0.870	0.620	0.780	0.710	0.750
BERT0p	0.742	0.941	0.958	0.667	0.836	0.781	0.813
BERT3p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
BERT6p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
ERNIE0p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
ERNIE3p	0.852	0.952	0.958	0.833	0.902	0.889	**0.896**

used less *um* than the control group. This indicates that unlike *uh*, *um* is more than a hesitation marker. Previous studies have also reported that children with autism spectrum disorder produced *um* less frequently than typically developed children (Gorman et al., 2016; Irvine et al., 2016), and that *um* was used less frequently during lying compared to truth-telling (Benus et al., 2006; Arciuli et al., 2010). All these results seem to suggest that *um* carries a lexical status and is retrieved in speech production. One possibility is that people with AD or autism have difficulty in retrieving the word *um* whereas people who are lying try not to use this word. More research is needed to test this hypothesis.

From our results, encoding pauses in the input was helpful for both BERT and ERINE fine-tuning for the task of AD classification. Pauses are ubiquitous in spoken language. They are distributed differently in fluent, normally disfluent, and abnormally disfluent speech. As we can see from **Figure 2**, the group with AD used more pauses and especially more long pauses than the control group. With pauses present in the text, the self-attention mechanism in BERT and ERNIE may learn how the pauses are correlated with other words, for example, whether there is a long pause between the determiner *the* and the following noun, which occurs more frequently in AD speech. We think this is part of the reason why encoding pauses improved the accuracy. There was no difference between three pauses and six pauses in terms of improvement in accuracy. More studies are needed to investigate the categories of pause length and determine the optimal number of pauses to be encoded for AD classification.

ERNIE was designed to learn language representation enhanced by knowledge masking strategies, including entity-level masking and phrase-level masking. Through these strategies, ERNIE "implicitly learned the information about knowledge and longer semantic dependency, such as the relationship between entities, the property of a entity and the type of a event". (Sun et al., 2019) We think this may be why ERNIE performs better on recognition of Alzheimer's speech, in which memory loss causes not only language problems but also difficulties of recognizing entities and events.

Both BERT and ERNIE were pre-trained on text corpora, with no pause information. Our study suggests that it may be useful to pre-train a language model using speech transcripts (either solely or combined with text corpora) that include pause information.

6 CONCLUSION

Accuracy of 89.6% was achieved on the test set of the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge, with ERNIE fine-tuning, plus an encoding of pauses. There is a high variance in BERT and ERNIE fine-tuning on a small training set. Our proposed ensemble method improves the accuracy and reduces variance in model performance. Pauses are useful in BERT and ERNIE fine-tuning for AD classification. *um* was used much less frequently in AD, suggesting that it may have a lexical status.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JY, Principal investigator, corresponding author; XC, Running ERNIE experiments; YB, Help running BERT experiments; ZY, Consultation on Alzheimer's disease, paper editing and proofreading; KC, Visualization of LOO experiment results, paper editing and proofreading.

ACKNOWLEDGMENTS

We thank Julia Li and Hao Tian for their suggestion and help with ERNIE. This paper is an extended version of our paper presented at Interspeech 2020 (Yuan et al., 2020).

REFERENCES

Arciuli, J., Mallard, D., and Villar, G. (2010). "Um, i can tell you're lying": linguistic markers of deception versus truth-telling in speech. *Appl. Psycholinguist.* 31, 397–411. 10.1017/S0142716410000044

Benus, S., Enos, F., Hirschberg, J., and Shriberg, E. (2006). "Pauses in deceptive speech," in Speech prosody 2006, Dresden, Germany, May 2–5, 2006.

Brown, E., and Miron, M. (1971). Lexical and syntactic predictors of the distribution of pause time in reading. *J. Verb. Learn. Verb. Behav.* 10, 658–667. doi:10.1016/S0022-5371(71)80072-5

Butcher, A. (1981). *Aspects of the speech pause: phonetic correlates and communicative functions*. Kiel, Germany: Institut fur Phonetik der Universitat Kiel.

Clark, H. H., and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition* 84, 73–111. doi:10.1016/s0010-0277(02)00017-3

Clark, H. H. (2006). *Pauses and hesitations: psycholinguistic approach*. Encyclopedia of Language & Linguistics, 244–248.

Corley, M., and Stewart, O. (2008). Hesitation disfluencies in spontaneous speech: the meaning of um. *Language and Linguistics Compass* 2, 589–602. 10.1111/j.1749-818X.2008.00068.x

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *J. Psycholinguist. Res.* 20, 445–464. 10.1007/BF01067637

de Ipiña, K. L., de Lizarduy, U. M., Calvo, P. M., Beitia, B., Garcia-Melero, J., Ecay-Torres, M., et al. (2017). "Analysis of disfluencies for automatic detection of mild cognitive impartment: a deep learning approach," in International Conference and Workshop on Bioinspired Intelligence (IWOBI), 2017, 1–4.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. Available at: https://arxiv.org/abs/1810.04805 (Accessed October 11, 2018).

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. (2020). Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. arXiv preprint. Available at: https://arxiv.org/abs/2002.06305 (Accessed February 15, 2020).

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *J. Mem. Lang.* 30, 210–233.

Filiou, R.-P., Bier, N., Slegers, A., Houzé, B., Belchior, P., and Brambati, S. M. (2019). Connected speech assessment in the early detection of alzheimer's disease and mild cognitive impairment: a scoping review. *Aphasiology.* 34, 1–33. 10.1080/02687038.2019.1608502

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *J Alzheimers Dis.* 49 (2), 407–422. doi:10.3233/JAD-150520

Fritsch, J., Wankerl, S., and Nöth, E. (2019). "Automatic diagnosis of alzheimer's disease using neural network language models," in ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing, Brighton, United Kingdom, May 12, 2020 (ICASSP IEEE), 5841–5845.

Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Lang. Speech* 4, 232–237. doi:10.1177/002383096100400405

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *Boston diagnostic Aphasia examination*. 3rd Edition. Philadelphia: Lippincott Williams & Wilkins.

Gorman, K., Olson, L., Hill, A., Lunsford, R., Heeman, P., and van Santen, J. (2016). Uh and um in children with autism spectrum disorders or language impairment. *Autism Res.* 9, 854–865. doi:10.1002/aur.1578

Gosztolya, G., Vincze, V., Toth, L., Pakaski, M., Kalman, J., and Hoffmann, I. (2019). Identifying mild cognitive impairment and mild alzheimer's diseasebased on spontaneous speech using asr and linguistic features. *Comput. Speech Lang.* 53, 181–197. 10.1016/j.csl.2018.07.007

Grosjean, F., Grosjean, L., and Lane, H. (1971). The patterns of silence: performance structures in sentence production. *Cognit. Psychol.* 11, 58–81. doi:10.1016/0010-0285(79)90004-5

Hawthorne, K., and Gerken, L. (2014). From pauses to clauses: prosody facilitates learning of syntactic constituency. *Cognition* 133, 420–428. doi:10.1016/j.cognition.2014.07.013

Irvine, C. A., Eigsti, I. M., and Fein, D. (2016). Uh, um, and autism: filler disfluencies as pragmatic markers in adolescents with optimal outcomes from autism spectrum disorder. *J. Autism Dev. Disord.* 46, 1061–1070. doi:10.1007/s10803-015-2651-y

Krivokapic, J. (2007). Prosodic planning: effects of phrasal length and complexity on pause duration. *J. Phonetics* 35, 162–179. doi:10.1016/j.wocn.2006.04.001

Laske, C., Sohrabi, H. R., Frost, S., López-de-Ipiña, K., Garrard, P., Buscema, M., et al. 2015). Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimers Dement* 11, 561–578. doi:10.1016/j.jalz.2014.06.004

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge," in Proceedings of INTERSPEECH 2020, Shanghai, China, October 25–29, 2020.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Mattson, M. P. (2004). Pathways towards and away from Alzheimer's disease. *Nature* 430, 631–639. doi:10.1038/nature02621

Mueller, K. D., Koscik, R. L., Hermann, B., Johnson, S. C., and Turkstra, L. S. (2017). Declines in connected language are associated with very early mild cognitive impairment: results from the Wisconsin registry for alzheimer's prevention. *Front. Aging Neurosci.* 9, 437. doi:10.3389/fnagi.2017.00437

Palo, F. D., and Parde, N. (2019). "Enriching neural models with targeted features for dementia detection," in Proceedings of the 57th annual Meeting of the Association for computational linguistics (ACL), Florence, Italy, July 2019.

Pulido, M. L. B., Hernández, J. B. A., Ballester, M. A. F., González, C., Mekyska, J., and Smékal, Z. (2020). Alzheimer's disease and automatic speech analysis: a review. *Expert Syst. Appl.* 150, 113213. doi:10.1016/j.eswa.2020.113213

Ramanarayanan, V., Goldstein, L., Byrd, D., and Narayanan, S. (2013). An investigation of articulatory setting using real-time magnetic resonance imaging. *J. Acoust. Soc. Am.* 134, 510–519. doi:10.1121/1.4807639

Ramig, L., Countryman, S., Thompson, L., and Horii, Y. (1995). Comparison of two forms of intensive speech treatment for Parkinson disease. *J. Speech Hear. Res.* 38, 1232–1251. doi:10.1044/jshr.3806.1232

Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *J. Psycholinguist. Res.* 2, 51–81. doi:10.1007/BF01067111

Schepman, A., and Rodway, P. (2000). Prosody and parsing in coordination structures. *Q. J. Exp. Psychol.* 53, 377–396. doi:10.1080/713755895

Shea, C., and Leonard, K. (2019). Evaluating measures of pausing for second language fluency research. *Can. Mod. Lang. Rev.* 75, 1–20. 10.3138/cmlr.2018-0258

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., et al. (2019). Ernie 2.0: a continual pre-training framework for language understanding. arXiv preprint. Available at: https://arxiv.org/abs/1907.12412 (Accessed July 29, 2019).

Tottie, G. (2011). Uh and um as sociolinguistic markers in british English. *Int. J. Corpus Linguist.* 16, 173–197. 10.1075/ijcl.16.2.02tot

Tran, T., Toshniwal, S., Bansal, M., Gimpel, K., Livescu, K., and Ostendorf, M. (2018). Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. arXiv preprint. Available at: https://arxiv.org/abs/1704.07287 (Accessed April 24, 2017).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in neural information processing systems*, 5998–6008.

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., and Liberman, M. (2016). Variation and change in the use of hesitation markers in germanic languages. *Lang. Dynam. Change* 6, 199–234. 10.1163/22105832-00602001

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," in Proceedings of INTERSPEECH 2020, Shanghai, China, October 25–29, 2020.

Yuan, J., and Liberman, M. (2008). Speaker identification on the scotus corpus. *J. Acoust. Soc. Am.* 123, 3878. doi:10.1121/1.2935783

Yuan, J., Xu, X., Lai, W., and Liberman, M. (2016). Pauses and pause fillers in Mandarin monologue speech: the effects of sex and proficiency. *Proc. Speech Prosody* 2016, 1167–1170. doi:10.21437/SpeechProsody.2016-240

Zellner, B. (1994). "Pauses and the temporal structure of speech," in *Fundamentals of speech synthesis and speech recognition*. Editor E. Keller (Chichester: John Wiley), 41–62.

Conflict of Interest: Authors JY, XC, YB and KC were employed by company Baidu USA Inc.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 05 February 2021
doi: 10.3389/fnagi.2021.623607

Acoustic and Language Based Deep Learning Approaches for Alzheimer's Dementia Detection From Spontaneous Speech

Pranav Mahajan[1]* *and Veeky Baths*[2]*

[1] *Cognitive Neuroscience Lab, Department of Electrical and Electronics Engineering, BITS Pilani University K. K. Birla Goa Campus, Pilani, India,* [2] *Cognitive Neuroscience Lab, Department of Biological Sciences, BITS Pilani University K. K. Birla Goa Campus, Pilani, India*

Current methods for early diagnosis of Alzheimer's Dementia include structured questionnaires, structured interviews, and various cognitive tests. Language difficulties are a major problem in dementia as linguistic skills break down. Current methods do not provide robust tools to capture the true nature of language deficits in spontaneous speech. Early detection of Alzheimer's Dementia (AD) from spontaneous speech overcomes the limitations of earlier approaches as it is less time consuming, can be done at home, and is relatively inexpensive. In this work, we re-implement the existing NLP methods, which used CNN-LSTM architectures and targeted features from conversational transcripts. Our work sheds light on why the accuracy of these models drops to 72.92% on the ADReSS dataset, whereas, they gave state of the art results on the DementiaBank dataset. Further, we build upon these language input-based recurrent neural networks by devising an end-to-end deep learning-based solution that performs a binary classification of Alzheimer's Dementia from the spontaneous speech of the patients. We utilize the ADReSS dataset for all our implementations and explore the deep learning-based methods of combining acoustic features into a common vector using recurrent units. Our approach of combining acoustic features using the Speech-GRU improves the accuracy by 2% in comparison to acoustic baselines. When further enriched by targeted features, the Speech-GRU performs better than acoustic baselines by 6.25%. We propose a bi-modal approach for AD classification and discuss the merits and opportunities of our approach.

Keywords: affective computing, cognitive decline detection, natural language processing, deep learning, computational paralinguistics

OPEN ACCESS

Edited by:
Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:
Poorya Parvizi,
University of Edinburgh,
United Kingdom
Masashi Kameyama,
Tokyo Metropolitan Geriatric Hospital
and Institute of Gerontology
(TMGH-IG), Japan

Correspondence:
Pranav Mahajan
f20170277@goa.bits-pilani.ac.in
Veeky Baths
veeky@goa.bits-pilani.ac.in

Received: *30 October 2020*
Accepted: *12 January 2021*
Published: *05 February 2021*

Citation:
Mahajan P and Baths V (2021)
Acoustic and Language Based Deep
Learning Approaches for Alzheimer's
Dementia Detection From
Spontaneous Speech.
Front. Aging Neurosci. 13:623607.
doi: 10.3389/fnagi.2021.623607

1. INTRODUCTION

Alzheimer's disease and related dementia disorders constitute a significant cause of disability and dependency among older adults worldwide and are among the costliest diseases in society. By 2030, it is estimated that the global cost of dementia could grow to US$ 2 trillion, which could overwhelm health and social care systems (Wimo et al., 2017). Alzheimer's Dementia (AD) is an irreversible brain disease that results in a gradual decrease in an individual's cognitive functioning. The main risk factor for AD is age, and therefore its highest incidence is amongst the elderly. However, if detected early, we can slow down or halt the degeneration with appropriate medication. Current methods of diagnosis usually involve lengthy medical evaluations, including lengthy questionnaires. There is an urgency for cost-efficient and scalable methods that can identify AD from an early stage. Thus, researchers worldwide are trying to find non-invasive early detection methods and treatments for these disorders.

Early symptoms of dementia are characterized by difficulty in word-finding, impaired reasoning, changes in language and speech, etc. This makes current research methodologies in speech and language processing suitable to be applied for early detection of cognitive impairment and AD. AD detection from spontaneous speech has been approached using speech input-based methods, language-based (text input-based) methods, and multi-modal approaches. Deep learning is a part of a broader family of machine learning methods based on artificial neural networks with representation learning. In prior work using language-based methods, we observe that deep learning based approaches (Orimaye et al., 2016; Karlekar et al., 2018; Di Palo and Parde, 2019; Kong et al., 2019) outperform pre-deep learning approaches (Orimaye et al., 2014; Fraser et al., 2016) on the DementiaBank dataset (Becker et al., 1994). Motivated by the shortcomings of manual feature-engineering for such a diverse and complex task, Karlekar et al. (2018) propose deep learning models—Convolutional neural network (CNN), Long short-term memory network (LSTM), and CNN-LSTM, to detect AD using just the conversational transcripts with minimal feature engineering using just word embeddings and parts-of-speech (POS) tags. Word embedding is any set of language modeling where words from a vocabulary are mapped to a vector of real numbers. POS-tagging is assigning a parts-of-speech tag to every word in the corpus, depending on it's context and definition. It is a method of enriching the feature processing stream. CNN layers are locally connected layers and pick up features in shorter time windows, where as LSTM layer is a type of recurrent neural network (RNN) layer which learns features and remembers features over longer timesteps. Recurrent layer or recurrent unit is any layer whose output not only depends on the input at the current timestep but also it's hidden state in the previous timestep. Thus, a CNN-LSTM architecture uses convolutional layers early on for feature extraction and then LSTM layers to learn patterns in a sequence. Di Palo and Parde (2019) further enrich the deep neural network models by Karlekar et al. (2018) by using targetted psycholinguistic, sentiment, and demographic features and also use class weight correction to handle class imbalance in the DementiaBank dataset (Becker et al., 1994). We build upon the work by Karlekar et al. (2018) and Di Palo and Parde (2019) and extend to multi-modal inputs and address the challenges that come with effectively combining features from multiple modalities for AD detection.

Amongst speech input-based methods, prior work has been more focused on using handcrafted acoustic features (Beltrami et al., 2016; Ambrosini et al., 2019) such as pitch, unvoiced duration, shimmer, pause duration, speech rate, or using feature banks. Haider et al. (2019) and Luz et al. (2020) use feature banks such as such as emobase, eGeMAPS (Eyben et al., 2015), ComParE (Eyben et al., 2013), and MRCG functionals (Chen et al., 2014) for feature extraction from speech segments. These features are not necessarily designed specifically for AD speech but capture various paralinguistic features relevant to AD speech. Effectively combining these features from various speech segments is an ongoing research problem that our work addresses. Previously, Haider et al. (2019) address it by proposing a new Active Data Representation method (ADR) to combine the features from a variable number of recordings into a fixed dimensional feature vector. They get the best results using the eGeMAPS feature set and even better results using a hard fusion of the feature sets. However, these methods fail to capture the temporal dynamics across the segments to the full extent. In this work, by using a recurrent unit, we combine the speech segment features in a fixed dimension vector while learning the features across the time span of the participant's conversation session. Chien et al. (2019) implement a bidirectional RNN on speech features extracted using a feature bank and propose an end-to-end method for automatic assessment of cognitive decline, but are restricted to speech input and do not extend to multi-modal inputs. Amongst multi-modal approaches using spontaneous speech, Zargarbashi and Babaali (2019) propose a model that extracts a perplexity score from the transcripts using an N-gram model extract I-vectors and X-vectors from the speech input. The concatenation of these feature vectors is then passed on to an SVM for AD classification. X-vectors and I-vectors are speech embeddings used in speaker recognition tasks, especially with speech segments of variable lengths. They use these embeddings even though AD diagnosis and speaker recognition are different tasks, as the voice biometrics and Alzheimer's signs are similar to an extent as both need to extract some specific patterns from captured signal contaminated with variations from various irrelevant sources. This prior work mentioned is relevant to our work because our work focuses on some of the open research problems, such as—How to capture complex patterns and temporal relations in speech and language modalities? And more importantly are there temporal patterns in the acoustic features extracted using the feature sets mentioned above, which can prove to be useful early detection of AD.

The majority of the previous results have been benchmarked on subsets from the Cookie theft task from the DementiaBank dataset (Becker et al., 1994) except the work by Chien et al. (2019) where they use NTUH Dataset which is a combination of multiple datasets such as Mandarin_Lu dataset (MacWhinney et al., 2011), NTU dataset (Chien et al., 2019), and 20 more participants from independently collected data. Dementia Bank dataset includes multiple transcripts from the same participant and has a significant imbalance in the age and gender distribution of the participants. ADReSS dataset (Luz et al., 2020) tries to mitigate these issues, and thus we use the ADReSS dataset in our work.

In this work, we address this by proposing a network that can train on speech segments using recurrent units and can be integrated with existing language-based deep learning models, which can also be enriched with targeted features.

Our contributions are as follows:

1. We re-implement the prior work by Karlekar et al. (2018) and Di Palo and Parde (2019) and benchmark the results on the new shared standardized ADReSS dataset.

2. We explore the deep learning-based methods of combining acoustic features into a common vector using recurrent units and propose a bi-modal approach for both the tasks.
3. We discuss the possibilities of further enriching the acoustic processing stream using features specific to AD speech and propose a bi-modal model based on concatenation of latent outputs of acoustic and language based models.

2. MATERIALS AND METHODS

2.1. ADReSS Dataset

Most earlier methods use a subset of the DementiaBank (Becker et al., 1994). Cookie theft task provides the largest source of unstructured speech and text data and thus has been used in Karlekar et al. (2018), Di Palo and Parde (2019), and Kong et al. (2019). The subset used in Di Palo and Parde (2019) includes multiple transcripts from the same participants, thus comprises a total of 243 transcripts from 104 non-AD participants and 1,049 transcripts from 208 AD participants. It also has imbalances in age and gender distribution. ADReSS Challenge dataset (Luz et al., 2020) tries to mitigate these issues. ADReSS Challenge dataset includes one full-wave audio (one session) per subject with accompanying conversational transcript. It also has a balanced distribution in terms of classes, age, and gender. As a result, we notice more than ten times reduction in dataset size in terms of the number of transcripts or full-wave session audios when compared to the dataset used in Karlekar et al. (2018), Di Palo and Parde (2019), and Kong et al. (2019). This is important to us since deep learning methods proposed in the previously mentioned approaches require larger amounts of data, reduction in data size, and removal of imbalance in the dataset can significantly affect replicability of results. ADReSS Challenge dataset includes data from 82 AD and 82 non-AD participants, of which 54 AD and 54 non-AD participants are included in the train set. The full-wave audio from each participant is further divided into an average of 24.86 (standard deviation $sd = 12.84$) normalized speech segments per participant using voice activity detection.

2.2. Classification Models and Approach

In this section, we'll briefly explain the language-based (transcript text input), acoustic feature-based and bi-modal models that we propose and progressively build on.

2.2.1. Language-Based Models

We first implement a CNN-LSTM model (Model A0) as proposed in Karlekar et al. (2018), which takes word embeddings (GloVe) as well as POS-tags as input, through two input streams, finally concatenated in a dense layer before passing it to the output layer. A dropout rate of 0.5 was used between the CNN and LSTM layer to prevent overfitting. We then implement the Model A1, as proposed by Di Palo and Parde (2019). It improves upon Model A0 by replacing the unidirectional LSTM in Model A0 with bidirectional LSTM layers with the insertion of attention mechanism on the hidden states of the LSTM and by including a dense neural network at the end of the LSTM layer to include targeted psycholinguistic, sentiment, and demographic features as described in Di Palo and Parde (2019). These targeted features are further explained in section 2.3. For models A0 and A1, we don't need to implement class weighting as done in Di Palo and Parde (2019) as ADReSS dataset doesn't have a class imbalance. Schematic representation of Models A0 and A1 can be found in **Figure 1i**.

2.2.2. Acoustic-Feature Based Models

Similar to how previous models have proposed a recurrent unit based language processing stream which is later further enriched with targeted features, we propose a similar approach of using speech input stream and taking acoustic features into account, which is later enriched with relevant, targeted features. These acoustic features are extracted from audio segments. The Model B0 is comprised of a Speech-GRU, which is defined by a recurrent layer (GRU) which takes in audio segment features per from each speech segment while maintaining the temporal structure across segments as in the full-wave audio session. The goal of this GRU unit is to combine the features from the speech segments into a common vector while maintaining the temporal structure across segments. A schematic of the GRU cell is included in **Figure 1iii**. We also briefly experimented with the Model B0, by replacing the unidirectional GRU with bidirectional GRU layers with the insertion of attention mechanism on the hidden states of the GRU. But, since they do not improve the performance significantly, we continue with the Speech-GRU in our further study. In Model B1, we progressively build upon Model B0, by enriching the speech input processing stream with various AD specific features extracted from lengths of speech segments provided by voice activity detection (VAD) and disfluency and interventional features as well as idea density-based features from complete transcripts and full-wave audio. Schematic representation of Models B0 and B1 can be found in **Figure 1ii**.

2.2.3. Bi-Modal Model

The Model that we propose is a direct combination of Model A1 and Model B1. The dense outputs from these two input streams is then concatenated and then connected to the output layer using dense connections. We use all targeted features from both the models in Model C. Schematic representation of Model C can be found in **Figure 1iv**.

2.3. Feature Extraction

In this subsection, we'll explain the targeted features used in Model A1, the acoustic feature sets used in Model B0, B1, C and the targeted features used in Model B1 and C. The targeted features used in Model A1, are token-level psycholinguistic features, token-level sentiment features and demographic features as described in Di Palo and Parde (2019). Each of the token-level features was averaged across all tokens in the instance, allowing us to obtain a participant-level feature vector to be coupled with the participant-level demographic features. The psycholinguistic features include (1) Age of acquisition of words which is the age at which a particular word is usually learned by individuals, (2) Concreteness which is a measure of word's tangibility, (3) Familiarity which is a

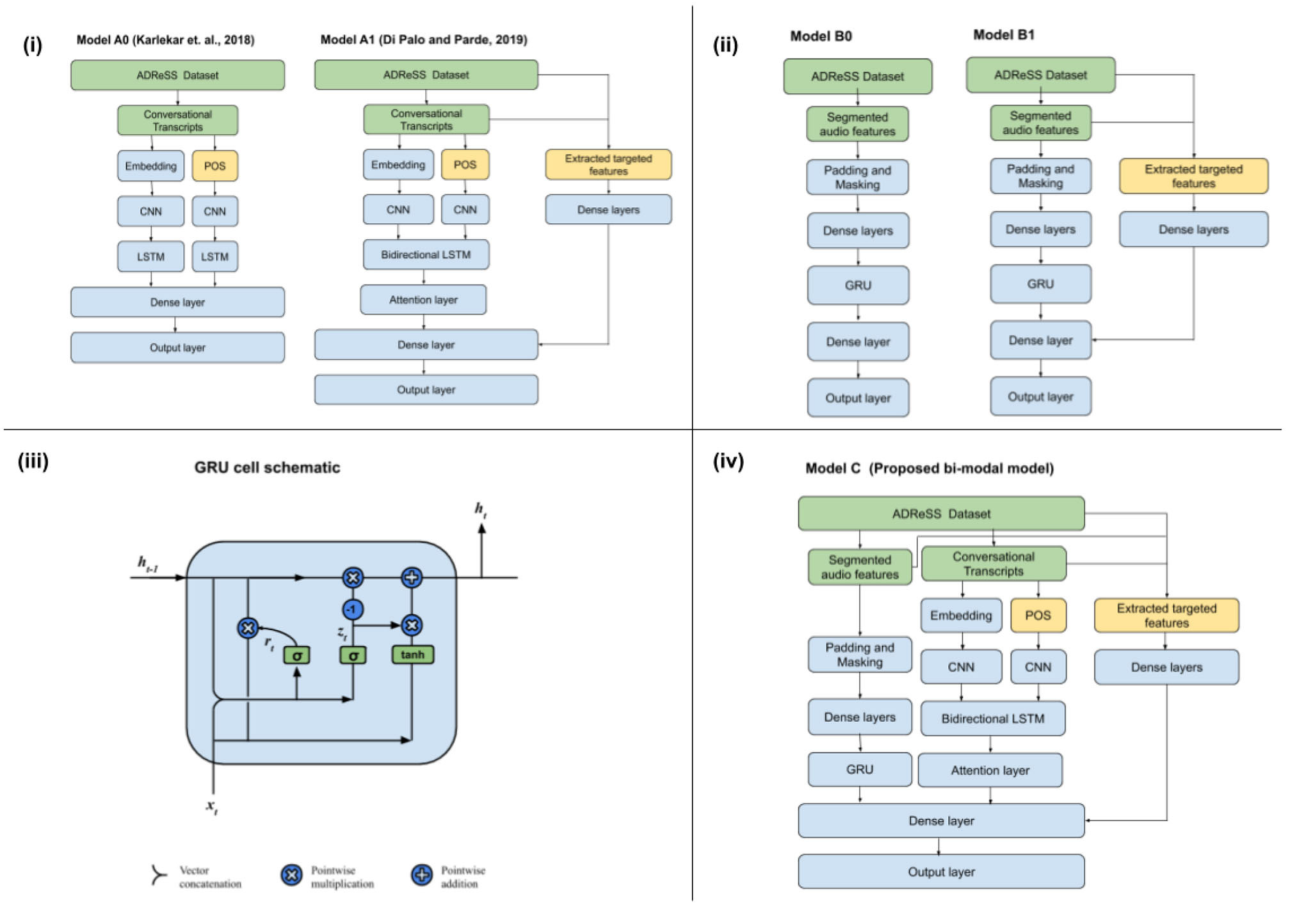


FIGURE 1 | (i) Language-based models, **(ii)** speech-based models, **(iii)** GRU cell schematic, and **(iv)** bi-modal model for AD detection.

measure of how often one might expect to encounter a word, (4) Imageability which is a measure of how easily a word can be visualized. Psycholinguistic features were obtained from an open-source repository[1] based on the work of Fraser et al. (2016). Sentiment scores were based around measuring the word's sentiment polarity and were obtained using the NLTK's sentiment library. The demographic features include participants age at the time of the visit and gender.

We compare the use of different feature banks for acoustic feature extraction, namely *emobase*, *eGeMAPS* (Eyben et al., 2015), and *ComParE* (Eyben et al., 2013) on Model B0 and then use the best performing feature set in Model B1 and C. These acoustic feature sets are described as follows.

emobase: This feature set (Schuller et al., 2010) contains the mel-frequency cepstral coefficients (MFCC) voice quality, fundamental frequency (F0), F0 envelope, line spectral pairs (LSP), and intensity features with their first and second-order derivatives. Several statistical functions are applied to these features, resulting in a total of 1,582 features for every speech segment. Haider et al. (2019) and Luz et al. (2020) use an older emobase feature set of 988 features, whereas we use the newer emobase2010 set from the INTERSPEECH 2010 Paralinguistics Challenge (Schuller et al., 2010).

eGeMAPS: The eGeMAPS feature set (Eyben et al., 2015) is a result of attempts to reduce other feature sets to a basic set of 88 features with theoretical significance (Eyben et al., 2015). The eGeMAPS features thus have the potential to detect physiological changes in voice production. It contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index, and slope V0 features, as well as their most common statistical functional.

ComParE: The ComParE feature set (Eyben et al., 2013) includes energy, spectral, MFCC, and voicing related low-level descriptors (LLDs). LLDs include logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity, and psycho-acoustic spectral sharpness. Statistical functionals are also computed, bringing the total to 6,373 features.

We used OpenSMILE[2] library for feature extraction using the emobase, eGeMAPS, ComParE feature bank. We performed

[1] https://github.com/vmasrani/dementia_classifier.

[2] https://www.audeering.com/opensmile/.

a Pearson correlation test on the whole dataset to remove acoustic features that were significantly correlated with the segment duration (when $R > 0.2$). Hence, 72 eGeMAPS, 1,072 emobase, and 3,056 ComParE features were not correlated with the duration of the speech chunks and were therefore selected for the machine learning experiments. The purpose of this step is to remove acoustic features correlated with the segment duration to remove the "local" features which are independent of segment duration while training the Model B0 purely on the low-level acoustic features. We later add global features such as mean, median, and standard deviation of all the segment lengths in an interview while training Model B1. Local features which are highly correlated with the segment duration can at times act as unnecessary noise and lead the machine learning models to learn spurious correlations. This preprocessing step is common with the approach by Luz et al. (2020) and Haider et al. (2019).

Our Model B1 is an extension of our Model B0, enriched with targetted features and our Model C is a combination of Model A1 and B1 and thus Model C uses targetted features from both models A1 and B1. The additional targetted features used in Model B1 and then subsequently in Model C are specific to AD speech and are obtained from a combination of speech segments, full wave audio as well as manually generated transcripts. These targetted features specific to AD speech can be broadly split into three categories—speech segment length-based features, disfluency, and interventional rate-based features and the features based on the concept of idea density. It is important to note that these features are not captured by our Model B0. Segment length features include six statistics about speech chunks segmented by the VAD. Disfluency and interventional features include a set of six distinct features from the transcripts, such word rate, intervention rate, and different kinds of pause rates reflecting upon speech impediments like slurring and stuttering, which show up in the transcripts in forms of "umm," "uhh" etc. Lastly, idea density based features comprise of the DEPID and DEPID-R features (Sirts et al., 2017) were computed as a measure of idea density. Idea density measures the rate at which ideas or elementary predications are expressed in an utterance or a text. Proportional idea density (PID) counts the expressed ideas and can be applied to any text. DEPID is a dependency-based method for computing PID and its version DEPID-R that enables to exclude repeating ideas which is a feature characteristic of AD speech.

2.4. Training and Validation Details

The following info is common to the training of all the models. We implement the models using Tensorflow 2.0 (Abadi et al., 2015). AdaGrad optimizer (Duchi et al., 2011) is used with a learning rate of 0.001. We train all the models for 200 epochs with early stopping as implemented in Di Palo and Parde (2019). All classification metrics use a classification threshold of 0.5.

The total dataset is split into a train dataset of 108 participants (54 AD and 54 non-AD participants) and test dataset of 48 participants (24 AD and 24 non-AD participants) as provided by Luz et al. (2020). Thus the test set is 30% of the total ADReSS dataset. K-fold cross validation (CV) is a useful CV strategy when sample size is lower as it uses every sample in the dataset but does not necessarily maintain balance in the labels (AD and non-AD) in each fold while splitting the train dataset into "k" folds. Performing a stratified k-fold CV assures this balance in labels in each fold and thus increases the reliability of metrics calculated on k-fold CV. We use 5-fold stratified cross-validation for all our models with the same seed value. We chose this cross-validation scheme over hold-out cross-validation schemes due to the small size of the dataset and to use every sample in the dataset. In Luz et al. (2020), the authors use leave one subject out (LOSO) cross-validation scheme, we find it infeasible in our case as training deep learning models are computationally more demanding and LOSO cross-validation scheme won't scale with more data without necessary compute requirements. For inference on test data, the models were trained on the complete train set for both the tasks separately and then tested on the test set.

3. RESULTS

The outputs of a binary classification algorithm fall into one of the four categories—true positives *tp*, false positives *fp*, false negatives *fn* and true negatives *tn*, depending on whether the predicted label matches with the true label or not. Recall is also known as Sensitivity or the true positive rate. Then classification metrics are defined as follows,

$$\text{Precision} = \frac{tp}{tp + fp} \tag{1}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{2}$$

$$\text{F1 score} = 2\frac{(\text{precision})(\text{recall})}{\text{precision} + \text{recall}} \tag{3}$$

In context of reproducing results by Karlekar et al. (2018) and Di Palo and Parde (2019) on ADReSS dataset, the classification task results (Precision, Recall, F1 score, and Accuracy) are shown in **Table 1** for 5-fold cross-validation and test setting, respectively. The results show that Model A1 performs better than Model A0 in all aspects of the classification task. We notice the difference between AD classification accuracy (0.8384 and 0.8820, respectively) achieved in Karlekar et al. (2018) and Di Palo and Parde (2019) on the complete Dementia Bank dataset and the AD classification accuracy achieved (0.6875 and 0.7292, respectively) by re-implementing those methods on ADReSS dataset.

In the context of the proposed acoustic feature processing Speech-GRU, the classification task results with the use of different acoustic feature set are shown in **Table 2** for 5-fold cross-validation and test sets, respectively. We observe that our model B0 with use of emobase as the acoustic feature set performs best followed by eGeMAPS and we observe that our recurrent model with ComParE features as input fails to learn. Our model B0 with the feature set emobase performs better than the acoustic feature-based baseline accuracy of 0.62 set by Luz et al. (2020). We use the best performing feature set (emobase) further, for our models B1 and C. We further also experimented with Speech-GRU in model B0 (emobase feature set) by replacing GRU layer

TABLE 1 | Validation and Test results of the language based models on the classification task.

Model	Val/Test	Class	Recall	Precision	F1 score	Accuracy
A0 (Karlekar et al., 2018)	5-fold CV	Non-AD	0.811 ± 0.085	0.637 ± 0.071	0.710 ± 0.059	0.673 ± 0.065
		AD	0.539 ± 0.121	0.752 ± 0.081	0.619 ± 0.090	
	Test set	Non-AD	0.8333	0.6451	0.7272	0.6875
		AD	0.5416	0.7647	0.6341	
A1 (Di Palo and Parde, 2019)	5-fold CV	Non-AD	0.836 ± 0.202	0.706 ± 0.152	0.735 ± 0.072	**0.710 ± 0.067**
		AD	0.600 ± 0.241	0.866 ± 0.167	0.654 ± 0.113	
	Test set	Non-AD	0.9167	0.6667	0.7719	**0.7292**
		AD	0.5416	0.8667	0.6667	

Bold values represent the validation and test accuracies of best performing model amongst the models under consideration in the respective table.

TABLE 2 | Validation and Test results of the Model B0 with different feature sets on the classification task.

Feature set	Val/Test	Class	Recall	Precision	F1 score	Accuracy
eGeMAPS	5-fold CV	Non-AD	0.527 ± 0.120	0.710 ± 0.151	0.581 ± 0.058	0.635 ± 0.034
		AD	0.745 ± 0.156	0.618 ± 0.029	0.667 ± 0.058	
	Test set	Non-AD	0.7500	0.5625	0.6428	0.5833
		AD	0.4166	0.6250	0.5	
emobase	5-fold CV	Non-AD	0.659 ± 0.094	0.704 ± 0.168	0.663 ± 0.057	**0.665 ± 0.082**
		AD	0.673 ± 0.219	0.664 ± 0.049	0.652 ± 0.125	
	Test set	Non-AD	0.6667	0.6400	0.6530	**0.6458**
		AD	0.6250	0.6521	0.6382	
ComParE	5-fold CV	Non-AD	0.441 ± 0.176	0.534 ± 0.139	0.475 ± 0.148	0.533 ± 0.129
		AD	0.625 ± 0.144	0.538 ± 0.132	0.573 ± 0.124	
	Test set	Non-AD	0.5833	0.5185	0.5490	0.5208
		AD	0.4583	0.5238	0.4888	

Bold values represent the validation and test accuracies of best performing model amongst the models under consideration in the respective table.

with a bidirectional GRU layer followed by the use of attention mechanism, but it resulted in validation accuracy of 0.6632 ± 0.0368 which did not significantly better than our basic Speech-GRU stream. Since we did not observe a significant improvement, we use our plain GRU stream for acoustic feature processing in models B1 and C.

The classification task results for the models B1 and C are shown in **Table 3**. Our results show that model B1, enriched with targeted features performs better than model B0 with an accuracy of 0.6875 on the test set. We further conduct ablation experiments on model B1 to tease out which of these targeted features contribute the most. The results of our ablation experiment in **Table 4** show that none of the targeted features (segment length based, disfluency, and interventional rate based and idea-density based) individually improve the test results of model B1, in comparison to model B0. But all of these features combined improve the classification accuracy of our model B1. Our model C benefits from linguistic feature processing stream of model A1 but does not perform better than model A1 in terms of test or validation accuracy. We notice a significant improvement in AD class Recall and a reduction in AD class Precision from model A1 to model C.

Finally, we include the Area under the Receiver-Operator characteristic curve for all the models in the **Figure 2** for quick comparison of the performance of all the models on the test set.

4. DISCUSSION

Amongst language-based models, the improvement in performance from model A0 to A1 can be attributed to the use of attention as well as the use of psycholinguistic and sentiment features. As per our results, model A0 and A1 which have shown the state of the art results on the complete Dementia bank dataset don't perform better than the linguistic feature baseline set by Luz et al. (2020) of accuracy 0.75 on the ADReSS dataset. This is important to note because the primary motivation of Karlekar et al. (2018) was to develop end to end deep learning method for AD detection with minimal feature engineering. Furthermore, noticing the difference in accuracy and F1 scores, there could be multiple factors involved in the success of Karlekar et al. (2018) and Di Palo and Parde (2019) and those that hinder the replicability of results on ADReSS dataset. The most prominent factor being, repeated occurrences

TABLE 3 | Validation and Test results of the Model B1 and Model C on the classification task.

Model	Val/Test	Class	Recall	Precision	F1 score	Accuracy
B1	5-fold CV	Non-AD	0.662 ± 0.175	0.670 ± 0.101	0.652 ± 0.125	0.662 ± 0.109
		AD	0.666 ± 0.170	0.675 ± 0.126	0.659 ± 0.139	
	Test set	Non-AD	0.8333	0.6452	0.7272	0.6875
		AD	0.5416	0.7647	0.6341	
C	5-fold CV	Non-AD	0.778 ± 0.104	0.673 ± 0.092	0.715 ± 0.070	**0.693 ± 0.082**
		AD	0.615 ± 0.151	0.743 ± 0.097	0.659 ± 0.112	
	Test set	Non-AD	0.8333	0.6896	0.7547	**0.7292**
		AD	0.6250	0.7894	0.6976	

Bold values represent the validation and test accuracies of best performing model amongst the models under consideration in the respective table.

of samples from the same participant in the Dementia Bank dataset. This could lead to significant overfitting to participant dependent features in models trained the DementiaBank dataset. As explained in section 2.1, DementiaBank has 243 transcripts from 104 non-AD participants whereas 1,049 transcripts from 208 AD participants. In comparison to that, ADReSS dataset includes only one transcript and full wave audio per participant, with 54 AD and 54 non-AD participants in the train set and 24 AD and 24 non-AD participants in the test set. Thus the total number of samples in DementiaBank is 1,292, which is around 8 times the dataset size of ADReSS. ADReSS dataset allows us to test the speaker independent nature of previously proposed models and our new model as there are no multiple sessions per participant. It is evident from other success of deep learning methods in other domains (not specific to AD speech) that such methods do scale with data, but that need not necessarily apply to tasks such as early detection of AD. Thus, we cannot take a purely minimal feature engineering approach, and future work should instead focus more on developing and utilizing features relevant to AD speech. Benchmarking on a dataset with more subjects in the future would help build a better understanding of whether these methods perform better compared to complete manual feature engineering-based solutions or not. Accuracy comparison of all the models with baselines on ADReSS dataset by Luz et al. (2020) as well as results on the DementiaBank dataset by Karlekar et al. (2018) and Di Palo and Parde (2019) can be found in **Figure 3**.

Our results from **Table 2** help us answer the question whether there exist temporal patterns relevant to AD detection in the acoustic features extracted using these feature sets emobase, eGeMAPs,ComParE etc. which are not explicitly designed for AD speech. Amongst the three feature sets, we observe that our Speech-GRU does pickup some relevant temporal patterns and effectively combines these features into a common vector. Our Speech-GRU with emobase feature set also performs better than the baseline by Luz et al. (2020), which takes the maximum vote of classification output of each of the speech segment. Still, the improvement is relatively small (2%). Moreover, the use of attention did improve the performance in language-based model A1, suggesting that there are temporal patterns which are relevant to AD speech in word vectors and POS-tags. But the same approach did not improve the performance in Speech-GRU, suggesting a general lack of temporal patterns across paralinguistic features of the speech segments. Future work could benefit from the development of AD specific feature sets.

TABLE 4 | Ablation experiments with Model B1 with different targeted features; Test results on classification task.

Targeted features	Class	Recall	Precision	F1 score	Accuracy
Seglen	Non-AD	0.5833	0.6363	0.6087	0.6250
	AD	0.6667	0.6154	0.6400	
Disf-inv	Non-AD	0.5000	0.6000	0.5454	0.5833
	AD	0.6667	0.5714	0.6154	
DEPID	non-AD	0.6250	0.6522	0.6383	0.6458
	AD	0.6667	0.6400	0.6530	
All combined	non-AD	0.8333	0.6452	0.7273	**0.6875**
	AD	0.5416	0.7647	0.6341	

Bold values represent the validation and test accuracies of best performing model amongst the models under consideration in the respective table.

It is important to note that our performance of model B0 is representative of the performance of AD detection without the use of any manual transcription. All the transcripts in Dementia Bank and ADReSS dataset are manually generated, and deploying this service would instead require automated transcription. Readers can refer to Zayats et al. (2019) for detailed analysis of impact of transcription errors (manual and automated) on automatic disfluency detection. Various disfluency and interventional features in our approach, as well as other state of the art approaches, rely on these manually generated transcripts for feature extraction and their performance may vary depending on whether the transcription is automated or not. In the ablation experiments, the decrease in the test accuracy in case of enrichment with disfluency and interventional features could be as these word rates, interventional rates, pause rates were extracted from manual transcripts. A better approach could be using forced alignment tools to get precise disfluency features, but since not all samples in the ADReSS dataset aligned with the transcript text, we didn't explore that idea further.

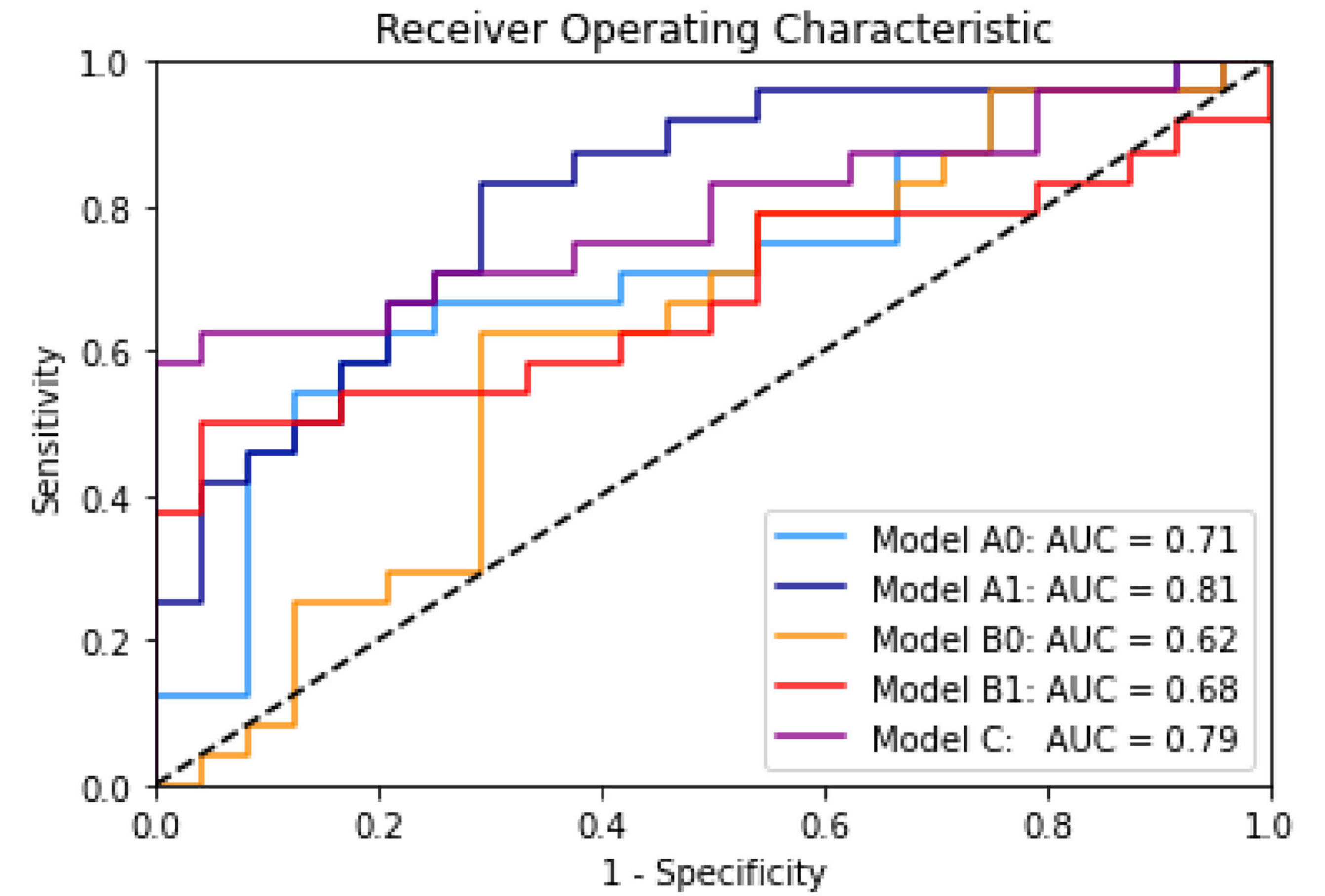


FIGURE 2 | Receiver operating characteristics for all Models A0, A1, B0, B1, and C and the area under the curve (AUC). Results on test set.

We observe that the language-based models A0 and A1 are characterized by higher non-AD class recall scores and higher AD class precision scores which are further aggravated from model A0 to A1. We observe that speech-based models were generally characterized by nearly equal precision and recall scores in AD and non-AD classes and we can also observe similar influence in the model C.

There are two possible reasons for the bimodal model C not performing significantly better than the language-based model A1, which are explained as follows. The first is that, the inherent representations learnt by the recurrent stream in Model A1 (trained on word embeddings and POS tags) and in Model B1 (on acoustic features of each segment, in lieu of acoustic embeddings) are quite different. And a mere concatenation of the final layers, can be thought of as a linear combination of the two representations and we observe that it does not provide rich space for a variety of cross-dimensional and non-linear combinations among the two representations. Because of this, the outputs of a Model B1 (which is a relatively weak learner in comparison to it's language counterpart Model A1) can act as noise in linear combination of these representations. This problem has been addressed by a variety of trainable feature aggregation methods, especially visual and language based representations, in the context of multimodal emotion detection or sentiment analysis. One of the most promising solution, which has proven to be successful in the context of multimodal sentiment analysis is focusing on word-level fusion (Chen et al., 2017), where they align the words with the speech segment of each word and generate combined Gated Multimodal Embeddings (GME), rather than combine the two representations in the final layers as we do in Model C. We believe a similar approach to generating combined word-level embeddings, where influence of each modality is also learnable through gating, can also help in the context of AD speech. Unfortunately, word-level fusion methods require alignment of both the modalities, which is very expensive in terms of reduction in data size as not all samples align even with the state of the art methods. Though this is a feasible option for other problems such as sentiment analysis, where data is in abundance and where the study can be carried out with a fraction of aligned data. But it's not a feasible option in small sized datasets like the ADReSS dataset as we observed while running the alignment tools (Montreal Forced Aligner[3]), <70% of the full wave audio samples aligned with the manually generated transcripts. The speech segment chunks provided by the ADReSS dataset use voice activity detection (VAD) and often include multiple words rather than providing a word-to-word alignment thus cannot be used for creating multimodal word embeddings. Readers can refer to Baltrušaitis et al. (2018) or a detailed survey of approaches and challenges faced in multi-modal machine learning in terms of representation, alignment, and fusion. Future work, in availability of more data, can attempt similar approaches to AD detection.

The second reason is that the idea density features used in Model B1 and then subsequently in Model C, have been computed using the transcripts. The disfluency and interventional rates used are also obtained from transcripts in lieu of aligning speech with transcripts. We compute the similarity in predictions of two models as ratio of predictions which match between two models upon total predictions in the test set (i.e., 48). We find the similarity between predictions of Model A1 and Model C to be 0.6667 whereas the similarity in

[3] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner.

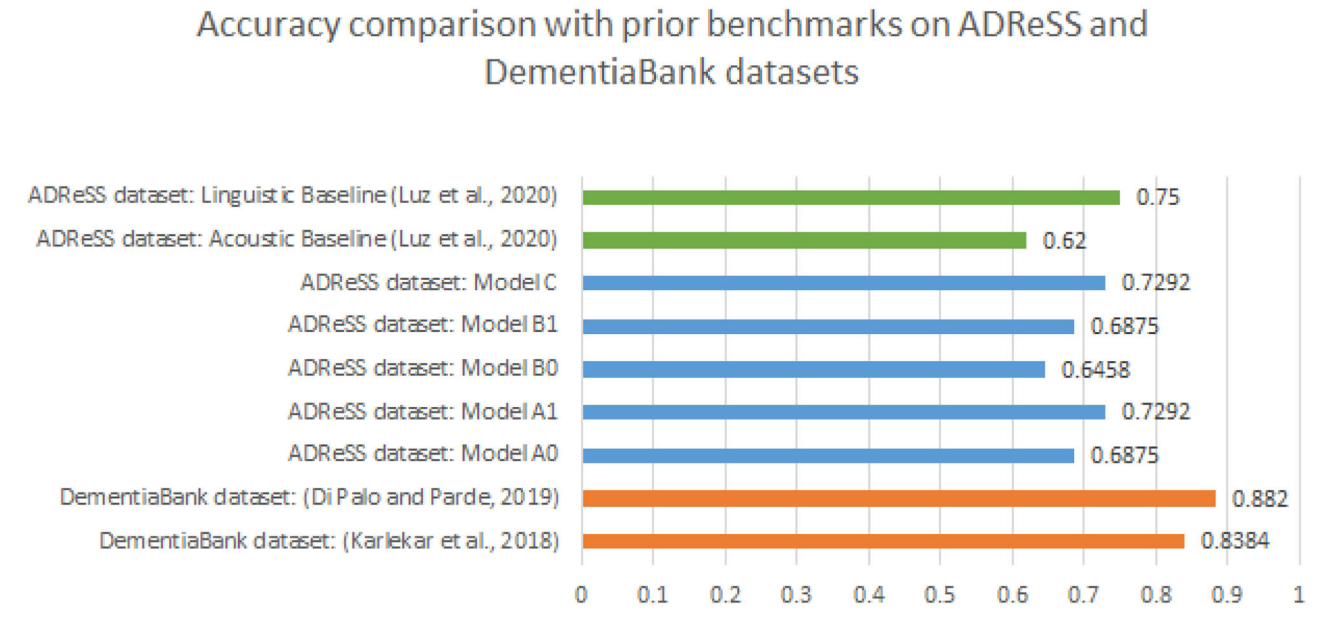
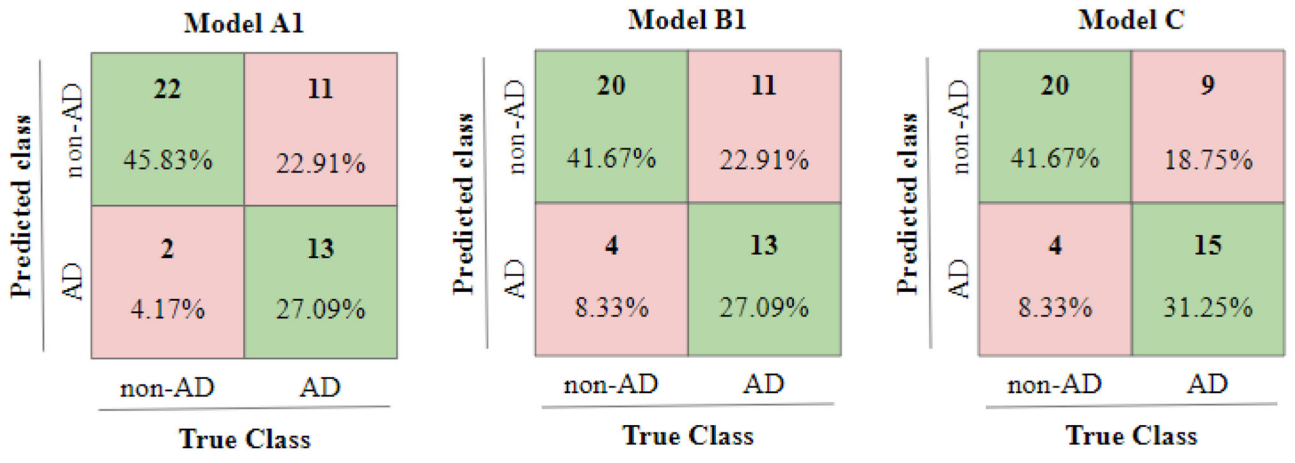


FIGURE 3 | Accuracy comparison of all Models A0, A1, B0, B1, and C with baselines on ADReSS dataset as well as referred state of the art approaches on DementiaBank dataset.


Model A1
Model B1
Model C
Predicted class
True Class

Model A1

Predicted class \ True Class	non-AD	AD
non-AD	22 45.83%	11 22.91%
AD	2 4.17%	13 27.09%

Model B1

Predicted class \ True Class	non-AD	AD
non-AD	20 41.67%	11 22.91%
AD	4 8.33%	13 27.09%

Model C

Predicted class \ True Class	non-AD	AD
non-AD	20 41.67%	9 18.75%
AD	4 8.33%	15 31.25%

FIGURE 4 | Confusion matrices of Test results of Model A1, B1, and C.

predictions of model B1 and model C to be 0.5416. Furthermore, we also observe that the similarity between predictions of Model A1 and B1 is 0.6667 and is greater than similarity between predictions of Model A1 and Model B0 which is 0.5834, suggesting that the additional targeted features obtained from transcripts and used in Model B1 might have already been captured in the Model A1 which trained only on the transcript data. Apart from the similarities in predictions, we can observe the confusion matrices of test predictions of Model A1, B1, and C in **Figure 4** which show the influence of Model A1 and B1 on Model C.

5. CONCLUSIONS

We re-implement existing deep learning-based methods on ADReSS dataset and discuss the challenges of the approach. We also introduce a bi-modal deep learning approach to AD classification from spontaneous speech and study in detail the Speech-GRU stream, which is further enriched with AD specific features through comprehensive comparisons of different variants. An important finding of this study is that the addition of targeted features increases the performance in AD detection in both language-based and acoustic-based models. Though the speech-GRU stream in our bi-modal approach is a relatively weaker learner compared to the language-based counterparts in the network, future work can aim at improving the acoustic feature extraction as well as a better combination of representations from different modalities. The Speech-GRU without and with extra targeted features performs much better than acoustic baselines and Model B0 is also representative of the extent of performance of solutions which don’t rely on manual transcription. Our results help us answer questions regarding the existence of temporal patterns relevant to AD detection in para-linguistic acoustic features often extracted using common feature sets as well as also address the reasons for a drop in accuracy of models on ADReSS dataset which were previously state of the art approaches on the complete Dementia Bank dataset.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://www.homepages.ed.ac.uk/sluzfil/ADReSS/.

AUTHOR CONTRIBUTIONS

PM participated in the ADReSS challenge, developed, and trained the machine learning models. VB helped in manuscript preparation and supervised the study. All authors contributed to the article and approved the submitted version.

FUNDING

The research project was funded by Cognitive Neuroscience Lab, BITS Pilani Goa which is funded by the Department of Science and Technology (DST) - DST-CSRI - SR/CSRI/50/2014(G), Govt. of India.

ACKNOWLEDGMENTS

We would like to thank the Electronics and Communications Engineering department, the Biological Sciences department and the Cognitive Neuroscience Lab at BITS Pilani University, K. K. Birla Goa Campus for their support and encouragement. We would like to thank the reviewers from the ADReSS challenge in Interspeech 2020 and the Frontiers research topic ADReSS for their very helpful feedback.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org

Ambrosini, E., Caielli, M., Milis, M., Loizou, C., Azzolino, D., Damanti, S., et al. (2019). “Automatic speech analysis to early detect functional cognitive decline in elderly population,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Berlin), 212–216. doi: 10.1109/EMBC.2019.8856768

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 423–443. doi: 10.1109/TPAMI.2018.2798607

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of alzheimer’s disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol.* 51, 585–594. doi: 10.1001/archneur.1994.00540180063015

Beltrami, D., Calzá, L., Gagliardi, G., Ghidoni, E., Marcello, N., Favretti, R. R., et al. (2016). “Automatic identification of mild cognitive impairment through the analysis of Italian spontaneous speech productions,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)* (Portorož), 2086–2093.

Chen, J., Wang, Y., and Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 1993–2002. doi: 10.1109/TASLP.2014.2359159

Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., Zadeh, A., and Morency, L.-P. (2017). “Multimodal sentiment analysis with word-level fusion and reinforcement learning,” in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow), 163–171. doi: 10.1145/3136755.3136801

Chien, Y.-W., Hong, S.-Y., Cheah, W.-T., Yao, L.-H., Chang, Y.-L., and Fu, L.-C. (2019). An automatic assessment system for Alzheimer’s disease based on speech using feature sequence generator and recurrent neural network. *Sci. Rep.* 9, 1–10. doi: 10.1038/s41598-019-56020-x

Di Palo, F., and Parde, N. (2019). Enriching neural models with targeted features for dementia detection. *arXiv [preprint]. arXiv1906.05483*. doi: 10.18653/v1/P19-2042

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12, 2121–2159. Available online at: https://jmlr.org/papers/v12/duchi11a.html

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2015). The Geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 190–202. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). "Recent developments in opensmile, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia* (New York, NY), 835–838. doi: 10.1145/2502081.2502224

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimer's Dis*. 49, 407–422. doi: 10.3233/JAD-150520

Haider, F., De La Fuente, S., and Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE J. Select. Top. Signal Process*. 14, 272–281. doi: 10.1109/JSTSP.2019.2955022

Karlekar, S., Niu, T., and Bansal, M. (2018). Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. *arXiv [preprint]. arXiv:1804.06440*. doi: 10.18653/v1/N18-2110

Kong, W., Jang, H., Carenini, G., and Field, T. (2019). "A neural model for predicting dementia from language," in *Machine Learning for Healthcare Conference* (Ann Arbor, MI), 270–286.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the address challenge. *arXiv [preprint]. arXiv:2004.06833*. doi: 10.21437/Interspeech.2020-2571

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: methods for studying discourse. *Aphasiology* 25, 1286–1307. doi: 10.1080/02687038.2011.589893

Orimaye, S. O., Wong, J. S.-M., and Fernandez, J. S. G. (2016). "Deep-deep neural network language models for predicting mild cognitive impairment," in *BAI@ IJCAI* (New York, NY), 14–20.

Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Quebec), 78–87. doi: 10.3115/v1/W14-3210

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., et al. (2010). "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association* (Makuhari).

Sirts, K., Piguet, O., and Johnson, M. (2017). Idea density for predicting Alzheimer's disease from transcribed speech. *arXiv [preprint]. arXiv:1706.04473*. doi: 10.18653/v1/K17-1033

Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., Prina, A. M., Winblad, B., et al. (2017). The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's Dement*. 13, 1–7. doi: 10.1016/j.jalz.2016.07.150

Zargarbashi, S., and Babaali, B. (2019). A multi-modal feature embedding approach to diagnose Alzheimer disease from spoken language. *arXiv [preprint]. arXiv:1910.00330*.

Zayats, V., Tran, T., Wright, R., Mansfield, C., and Ostendorf, M. (2019). Disfluencies and human speech transcription errors. Graz. *arXiv [preprint]. arXiv:1904.04398*. doi: 10.21437/Interspeech.2019-3134

Conflict of Interest: The 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.

BRIEF RESEARCH REPORT
published: 09 February 2021
doi: 10.3389/fcomp.2021.624659

Learning Language and Acoustic Models for Identifying Alzheimer's Dementia From Speech

Zehra Shah[1], Jeffrey Sawalha[2], Mashrura Tasnim[1], Shi-ang Qi[1], Eleni Stroulia[1] and Russell Greiner[1,2,3]*

[1]Department of Computing Science, University of Alberta, Edmonton, AB, Canada, [2]Department of Psychiatry, University of Alberta, Edmonton, AB, Canada, [3]Alberta Machine Intelligence Institute, Edmonton, AB, Canada

Alzheimer's dementia (AD) is a chronic neurodegenerative illness that manifests in a gradual decline of cognitive function. Early identification of AD is essential for managing the ensuing cognitive deficits, which may lead to a better prognostic outcome. Speech data can serve as a window into cognitive functioning and can be used to screen for early signs of AD. This paper describes methods for learning models using speech samples from the DementiaBank database, for identifying which subjects have Alzheimer's dementia. We consider two machine learning tasks: 1) binary classification to distinguish patients from healthy controls, and 2) regression to estimate each subject's Mini-Mental State Examination (MMSE) score. To develop models that can use acoustic and/or language features, we explore a variety of dimension reduction techniques, training algorithms, and fusion strategies. Our best performing classification model, using language features with dimension reduction and regularized logistic regression, achieves an accuracy of 85.4% on a held-out test set. On the regression task, a linear regression model trained on a reduced set of language features achieves a root mean square error (RMSE) of 5.62 on the test set. These results demonstrate the promise of using machine learning for detecting cognitive decline from speech in AD patients.

Keywords: speech and audio classification, pathological speech and language, automatic analysis of speaker states and traits, machine learning, natural language proceeding (NLP)

OPEN ACCESS

Edited by:
Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:
Marko Meža,
University of Ljubljana, Slovenia
Michal Przemyslaw Muszynski,
Université de Genève, Switzerland

****Correspondence:***
Zehra Shah
zshah@ualberta.ca

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *31 October 2020*
Accepted: *08 January 2021*
Published: *09 February 2021*

Citation:
Shah Z, Sawalha J, Tasnim M, Qi S, Stroulia E and Greiner R (2021) Learning Language and Acoustic Models for Identifying Alzheimer's Dementia From Speech. Front. Comput. Sci. 3:624659. doi: 10.3389/fcomp.2021.624659

1 INTRODUCTION

Alzheimer's Dementia (AD) has recently become one of the leading causes of death in people over 70 years (Alzheimer's Association (2019)). With life expectancy increasing, the prevalence of AD among older adults is also rising. Currently, the number of cases among people over the age of 60 is doubling every 4–5 years, and currently, one in every three individuals over the age of 80 is likely to develop AD (Ritchie and Lovestone (2002)). AD is a progressive neurodegenerative disorder that is characterized by the loss of subcortical neurons and synapses that begins in areas such as the hippocampus and the entorhinal cortex (Braak and Braak (1991); Terry et al., (1991)). Over time, more associative areas begin to show amyloid deposition and neurofibrillary tangles in addition to neuronal and synaptic loss. As it spreads, patients develop additional cognitive and functional deficits in domains such as attention, executive function, memory and language (Nestor et al., (2004)). Current theories maintain that clinical symptoms are preceded by subtle cognitive deficits that worsen over time. Early recognition of these deficits could prove valuable for treating pre-stage AD, allowing for a better quality of life for the patient and their caregivers.

Currently, clinical diagnostic methods for determining who has AD include cognitive assessments (e.g., Mini-Mental State Examination [MMSE]), self-report questionnaires and neuroimaging (e.g., Positron Emission Tomography [PET]) (Weller and Budson (2018)). While these methods have proven useful, they suffer from several shortcomings. Cognitive assessments can be tedious and suffer from low test-retest reliability based on practice effects; self-report questionnaires also lack reliability and validity; and neuroimaging is an expensive, invasive, and time-consuming procedure.

By contrast, speech analysis is a simple, non-invasive and inexpensive approach. There are several reasons why it may be useful for detecting AD. Early identification, especially in the prodromal stages, can significantly reduce the progression of various cognitive deficits (Dubois et al., (2009)). There is evidence that therapeutic interventions are most efficacious before neuronal degeneration occurs in the brain (Nestor et al., (2004)). Thus, an emphasis on early detection is imperative for the prognosis of AD. As such, episodic memory, visuospatial ability, and confusion are some of the first signs of cognitive decline in AD patients (Arnáiz and Almkvist (2003); Jacobs et al., (1995)). These deficits can be observed through verbal communication in a structured task, motivating the recent use of speech data for diagnostic screening of AD in elder patients (Chien et al., (2019)). In our study, we used machine learning (ML) approaches to distinguish between AD and control patients, using acoustic and linguistic features from spontaneous speech produced by a subject describing a picture.

The current literature on detecting AD from spontaneous speech samples can be divided into two main categories. One class of systems analyzes linguistic features (lexicon, syntactic and semantic information), while the other deals with acoustic-dependent features. In the acoustic domain, AD patients exhibit longer and more frequent hesitations, lower speech and articulation rates, and longer pauses compared to control participants in spontaneous speech tasks (Hoffmann et al., (2010); Szatloczki et al., (2015)). Some have attempted to apply ML approaches to learn models that use acoustic features to distinguish AD from control participants. Tóth et al., (2018) learned a model for distinguishing early stage AD patients from control patients using spontaneous speech from a recall task. Their classification model found significant differences in speech tempo, articulation rate, silent pause, and length of utterance. Mirzaei et al., (2017) tried to improve on previous models by examining temporal features (jitter, shimmer, harmonics-to-noize ratio, Mel frequency cepstral coefficients [MFCCs]).

Conversational transcripts contain rich information about the speaker, such as the wealth of their vocabulary, the complexity of their syntactic structures, and the information and meanings they communicate. Previous research has shown that language changes in patients who suffer from AD (Wankerl et al., (2017); Kempler (1995))–e.g., these patients often have difficulty naming objects within specific categories, replacing forgotten words with pronouns and repeating certain words or phrases (Kirshner (2012); Adlam et al., (2006); Nicholas et al., (1985)). This has motivated numerous research projects on conversation samples in AD and control patients. Fraser et al., (2016) examined picture description transcripts from demented vs. control individuals. Subsequently, they also analyzed acoustic features in addition to natural language, and achieved an accuracy of 81%. They found that semantic information was one of the best features (syntactic fluency, MFCCs and phonation rate) for separating AD from control participants.

Our paper is motivated by the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge, hosted by the INTERSPEECH 2020 conference (Luz et al., (2020)). The data set provided in this challenge is a carefully curated subset of the larger DementiaBank corpus (Becker et al., (1994)). Among the various challenge submissions, the top-performing models analyzed both linguistic and acoustic features, and many of these top submissions used deep learning methods (including some pre-trained models) to generate their results. For example, Koo et al., (2020) used an ensemble approach with bi-modal convolutional recurrent neural networks (cRNN), applied to a variety of feature sets from pre-trained acoustic and linguistic algorithms in addition to some hand-crafted features. They achieved an accuracy of 81.25% on their classifier evaluation and an RMSE score of 3.75. Another study by Balagopalan et al., (2020) achieved an accuracy of 83.33% and an RMSE of 4.56 by adding a binary classification layer to a pre-trained language algorithm developed by Google: Bidirectional Encoder Representations from Transformers–BERT. The Sarawgi et al., (2020) submission applied RNNs and multi-layered perceptrons (MLP) to various types of acoustic and linguistic features in an ensemble approach. They also used transfer learning from the classification models to the MMSE scores by modifying the last layer structure, achieving an RMSE of 4.6 and an accuracy of 83.33%. Lastly, Searle et al., (2020) used linguistic features only, with pre-trained Transformer based models, and achieved their best performance using features computed from the full transcripts (including both participant and interviewer speech). They obtained a classification accuracy of 81% and an RMSE of 4.58. The commonality among these top submissions was the use of deep-learning methods, along with pre-trained acoustic and/or language models.

Our study hopes to improve further by applying simple, computationally inexpensive ML techniques to natural language and acoustic information. In particular, we train models that use both acoustic and language features to distinguish AD from healthy age-matched elders and predict their MMSE scores. Our system feeds the acoustic features into one pipeline, and the linguistic ones in another. Each pipeline preprocesses the features, then uses internal cross-validation to tune the hyperparameters and select the relevant subset of features. We use ensemble methods to combine the various learned models, to produce models that can 1) label a speech sample as either AD or non-AD, and 2) predict the associated MMSE scores.

2 METHOD

For this study, we were given a training set of 54 AD patients and an age- and gender-matched set of 54 healthy controls (this is a subset of the larger DementiaBank data set; see Becker et al., (1994)). This subset of DementiaBank contained spontaneous speech samples of participants asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam (Goodglass et al., (2001)). For each participant, we obtained 1) the original recorded speech sample, 2) the normalized speech segments extracted from the full audio sample after voice activity detection, audio normalization and noise removal, as well as 3) the speech transcript files annotated using CHAT (Codes for Human Analysis of Transcripts) transcription format (MacWhinney (2017)). Additionally, some descriptive features were given about these individuals, including age, gender, binary class label (AD/non-AD; the target for the classification task), and their MMSE score (which we try to predict in the regression task). The MMSE has a maximum score of 30, and lower MMSE scores are generally associated with progressively more severe dementia. The challenge organizers withheld a test set containing data from 24 AD and 24 control participants for final evaluation. For further details of this data set, we refer the reader to Luz et al., (2020).

We considered a set of possible base learners, each over a subset of the features–the (1), (2), and (3) mentioned above. We used internal 5-fold cross validation to identify which of these base learners was best. Due to the size of our data, we chose to use a 5-fold CV procedure. 10-fold CV or Leave-one-out CV procedure would result in small partitions, leading to possible overfitting (low bias, higher variance). To ensure consistent and reliable comparison between our models, we defined and used a common set of folds that were balanced in terms of class labels (or MMSE scores) as well as gender. For each model, we evaluated performance metrics (average accuracy for classification, and average RMSE for regression) based on these test folds, as well as on the final hold-out test set.

2.1 Language and Fluency Features

The organizers provided transcripts that were annotated using the CHAT coding system (MacWhinney (2017)). First we extracted only the participant's speech from these transcripts (removing the interviewer's content). Then, using the CLAN (Computerized Language Analysis) program for processing transcripts in the CHAT format, we computed the following set of global syntactic and semantic features for each transcript: type-token ratio (TTR)–the number of unique words divided by total number of words; mean length of utterance (MLU), where an utterance is a speech fragment beginning and ending with a clear pause; number of verbs per utterance; percentage of occurrence of various parts of speech (nouns, verbs, conjunctions, etc.); number of retracings (self-corrections or changes); and number of repetitions. We also computed a number of fluency features, including percent of broken words, part-word and whole-word repetitions, sound prolongations, abandoned word choices, word and phrase repetitions, filled pauses, and non-filled pauses. In total, we computed 62 such informative summary features for each transcript.

2.2 N-Gram Features

We processed the raw (unannotated) transcripts to compute bag-of-words and bigram features. First, we standardized the transcripts by converting them into a list of word tokens. Next, we used the WordNet lemmatizer (Miller (1998)) to find and replace each word with the corresponding lemma; for example, words like "stands", "standing" and "stood" were all replaced by the common root word "stand". Finally, we removed stopwords from each transcript, where stopwords are highly common (and presumably uninformative) words that may add noise to the data (such as "I", "am", "was", etc.), using a predefined stopwords list from the Python natural language toolkit (NLTK) package.

Next, we used the standardized transcripts to compute bag-of-words vectors (using words seen in the training set only)–that is, a vector of 514 integers for each transcript, where the *k*th value is the number of times the *k*th word occurred–and normalized these vectors with the Term Frequency-Inverse Document Frequency (TF-IDF) function, which is a normalization procedure that reflects how important a word is to a document in a corpus–effectively penalizing words that occur frequently in most of the documents in the corpus. For example, in our case the word "boy" might occur frequently in all transcripts, so it may not be very informative. Finally, we also computed bigram vectors in a manner similar to bag-of-words–where each bigram is a *pair* of words that appeared adjacent to one another. We found a set of 2,810 bigrams.

2.3 Acoustic Features

Using the speaker timing information provided in the transcripts, we extracted the participants' utterances (removing the interviewer's voice) from the audio recordings, for a total of 1,501 participant utterances from the training set, and 592 from the test set. We then normalized the audio volume across all speech segments. We computed four different sets of features from each audio segment using OpenSMILE v2.1 (Eyben et al., (2010)). Note that our overall learner will consider various base-learners, each running on one of these feature sets.

(FeatureSet#1) The **AVEC 2013** (Valstar et al., (2013)) feature set includes 2,268 acoustic features including 76 low level descriptor (LLD) features and their statistical, regression and local minima/maxima related functionals. The LLD features include energy, spectral and voicing related features; delta coefficients of the energy/spectral features, delta coefficients of the voicing related LLDs and voiced/unvoiced duration based features.

(FeatureSet#2) The **ComParE 2013** (Schuller et al., (2013)) feature set includes energy, spectral, MFCC, and voicing related features, logarithmic harmonic-to-noize ratio (HNR), voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functionals are also computed, leading to a total of 6,373 features.

(FeatureSet#3) Our third feature set consists of the following three feature sets. The **emo_large** (Eyben et al., (2010)) feature set consists of cepstral, spectral, energy and voicing related features, their first and second order delta coefficients as LLDs; and their 39 statistical functionals. The functionals are computed over 20 ms frames in spoken utterances. This produced 6,552 acoustic features across the utterances. The **Jitter-shimmer**

feature set is a subset of INTERSPEECH 2010 Paralinguistic Challenge (Schuller et al., (2010)) feature set, consisting of three pitch related LLDs and their delta coefficients. We also computed 19 statistical functionals of the LLDs on the voiced sections of the utterances, resulting in 114 features. Finally, we extracted seven speech and articulation rate features by automatically detecting syllable nuclei (De Jong and Wempe (2009)), and used a script from the software program Praat to detect peaks in intensities (dB) followed by sharp dips. We also calculated other features, such as words per minute, number of syllables, phonation time, articulation rate, speech duration and number of pauses for each speech sample (Chakraborty et al., (2020)).

(FeatureSet#4) We computed the **MFCC 1–16** features and their delta coefficients from 26 Mel-bands, which uses the fast Fourier transform (FFT) power spectrum. The frequency range of the Mel-spectrum is set from 0 to 8 kHz. Inclusion of statistical functionals resulted in 592 features. This feature set is a subset of AVEC 2013 feature set (Valstar et al., (2013)).

We also added age and gender of the participants to each set of features.

2.4 Language-Based Models

Given our two sets of linguistic features above (**Sections 2.1 and 2.2**), we explored various dimension reduction techniques and base learning algorithms to find the best performing pipeline. The dimension reduction techniques include Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), and univariate feature selection using ANOVA F-values. The base learning algorithms explored for the classification task are logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB). For the regression task, the regression versions of the same algorithms are trained (except logistic regression is replaced by linear regression). Internal 5-fold cross-validation was used to tune the hyperparameters for each model based on accuracy. The hyperparameters explored were:

Dimension reduction: For classification models, dimension reduction with PCA using {10, 20, 30, 50} components, and LSA using {100, 200, 500} components; for regression models, dimension reduction with PCA using {20, 30, 50} components, and LSA using {200, 500, 800} components.

Models: SVM (regularization parameter C: {0.1, 1, 10, 100, 1,000}, kernel: {linear, RBF, polynomial}); LR (regularization parameter C: 20 values spaced evenly on a log scale in the range $[10^{-4}, 10^{4}]$, loss function: {L1, L2}); RF (number of trees: {100, 300, 500, 700}, maximum features at each split: {5, 15, 25, 35, 45, 55}, minimum samples at leaf node: {1, 2, 3, 4}); and XGB (maximum depth: {5, 6, 7, 8}, learning rate: {0.02, 0.05, 0.07, 0.1}, number of trees: {50, 100, 200, 500, 1,000}). The same hyperparameters were explored for the regression models as well (with the exception of replacing LR with linear regression).

Our internal cross-validation found the best-performing *language-based classification* model, which consisted of the following steps:

Step1: 5-component PCA transformation of the *dense* language and fluency features described in **Section 2.1** (after standardizing using z-scores);
Step2: 50-component LSA transformation of the *sparse* unigram and bigram features described in **Section 2.2** (after standardizing using TF-IDF transform); and
Step3: L1-regularized logistic regression.

The best language-based regression model involved the following:

Step1: 30-component PCA transformation of the *dense* language and fluency features described in **Section 2.1** (after standardizing using z-scores);
Step2: 100-component LSA transformation of the *sparse* unigram and bigram features described in **Section 2.2** (after standardizing using TF-IDF transform); and
Step3: Random Forest Regressor, using 100 trees, minimum of four instances at each leaf node, and 25 features considered for each split.

2.5 Acoustic Models

All acoustic features were real values and were therefore standardized using z-scores. We used PCA to reduce the dimensionality of the features sets. For FeatureSet#1 and FeatureSet#2, we used PCA, and kept the minimum number of features capable of retaining 95% of the variance. In case of FeatureSet#3 and FeatureSet#4, the number of principals were determined through internal 5-fold cross-validation. Therefore, the dimension of FeatureSet#1 is reduced from 2,268 to 700, FeatureSet#2 from 6,373 to 1,100, FeatureSet#3 from 6,552 to 1,000 and FeatureSet#4 from 592 to 50. Next, we selected the best 50 principal components from FeatureSet#1, and the best 70 from FeatureSet#3 applying univariate feature selection method based on ANOVA F-value between the label and each feature. For FeatureSet#2, we calculated feature importance weights using a decision-tree regression model, and selected only the features with importance weight higher than the mean.

After this pre-processing stage, our system fed these audio features to various machine-learning algorithms, that each identify patterns of features that can distinguish dementia patients from healthy controls (the classification task), and can compute a subject's MMSE score (the regression task). We explored several learning algorithms, including Adaboost, XGB, RF, gradient boosting (GBT), decision trees (DT), hidden Markov model (HMM) and neural network (NN). Internal 5-fold cross-validation was performed to tune the hyperparameters of the classifiers and regressors. The predictions were made in two steps. In the first step, the classifiers and regressors were trained and tested with acoustic features, age and gender to predict whether the speech segment was uttered by a health control or an AD patient and to predict that subject's MMSE score. Next, weighted majority vote classification was performed to assign each subject a label of health control or AD, based on the majority labels of the segment level classification. The predicted MMSE scores on all the segments of one subject were averaged to calculate the final MMSE score of that subject. The best performing classifiers on acoustic data are the following:

(1) Neural network with one hidden layer, trained on FeatureSet#1.

TABLE 1 | Results of our best performing classification models distinguishing AD from non-AD subjects. The "Baseline (Acoustic)" model is described in Luz et al. (2020). The right-most column shows accuracy on the held-out test set of 48 subjects (24 AD and 24 non-AD). The rest of the table lists model performance using 5-fold cross-validation on the training set of 108 subjects (54 AD and 54 non-AD).

Classifiers	Class	Precision	Recall	F1 score	Accuracy	Accuracy (hold-out set)
	AD	1.0	0.60	0.75		
Logistic regression (NLP)	HC	0.71	1.00	0.83	80% ± 0.00%	85%
	OVR	0.86	0.80	0.79		
	AD	0.68	0.84	0.75		
SVM (NLP)	HC	0.79	0.60	0.68	72% ± 1.85%	73%
	OVR	0.73	0.72	0.72		
	AD	0.74	0.96	0.83		
Majority vote (NLP + Acoustic)	HC	0.94	0.66	0.78	81% ± 1.17%	83%
	OVR	0.84	0.81	0.81		
	AD	0.71	0.78	0.74		
Majority vote (Acoustic)	HC	0.76	0.68	0.72	73% ± 1.36%	65%
	OVR	0.73	0.73	0.73		
	AD	0.57	0.52	0.54		
Baseline (Acoustic)	HC	0.56	0.61	0.58	57%	63%
	OVR	0.57	0.57	0.56		

AD, Alzheimer's dementia; HC, Healthy control; OVR, Overall rating.

(2) AdaBoost Classifier with 50 estimator and logistic regression as base estimator, trained on FeatureSet#4.
(3) Adaboost with 100 estimators and DT as the base estimator trained on FeatureSet#3.

The three regressors with the lowest RMSE were:

(1) Gradient boosting regressor, trained on FeatureSet#4.
(2) Decision tree with number of leaves 20, trained on FeatureSet#2.
(3) Adaboost regressor trained on FeatureSet#3 with 100 estimators.

2.6 Ensemble Methods

After obtaining our best-performing acoustic and language-based models, we computed a weighted majority-vote ensemble meta-algorithm for classification. We chose the three best-performing acoustic models along with the best-performing language model, and computed a final prediction by taking a linear weighted combination of the individual model predictions. The weights assigned to each model were proportional to that model's mean cross-validation accuracy, such that the best performing model is given the highest weight in the final prediction. For regression, we also computed an unweighted averaging of our best language and acoustic model predictions for MMSE scores.

3 RESULTS

3.1 Classification

Table 1 presents the results for the classification task. The model that obtained the highest average cross-validation accuracy (81% ± 1.17%) is a weighted-majority-vote ensemble of the best language-based model and three of the best acoustic-based models. The second highest accuracy (80% ± 0.00%) was obtained by the language-based logistic regression. However, a McNemar test reveals that these two models do not exhibit a statistically significant difference in performance (McNemar test statistic = 4.0, $p > 0.05$). This is also evident by the performance of these two models on the final held-out set, where the language-based logistic regression gives the highest accuracy (85%) and the weighted-majority-vote ensemble gives a slightly lower accuracy (83%). Using McNemar's test to compare these two models on the held-out test set, we obtain a test statistic of 3.0, with $p > 0.05$, indicating that the performance difference between these models is not statistically significant.

Note that our ensemble model, which uses only acoustic features, performs significantly better than the "baseline model" (provided by the organizers), which also uses acoustic features only.

3.2 MMSE Prediction

Table 2 shows the RMSE of various regression models; columns 2 and 3 show the average RMSE and R^2 scores over the five cross-validation folds, and columns 4 and 5, on the hold-out test set (provided by the organizers of the challenge). These results show that the language-based model obtains the best RMSE of 6.43 on the cross-validation set and 5.62 on the hold-out set. The combined language-acoustic model did not perform as well as the standalone language-based model, with an average RMSE of 6.83 on the cross-validation set and 6.12 on the hold-out set.

Further, the Wilcoxon test between the RMSEs of the two best models (best acoustic + best language-based combination vs. best stand-alone language-based), returns a test statistic of 66.0 with $p < 0.05$ on the hold-out set, and a test statistic of 1,375.0 with $p < 0.05$ on the cross-validation set. This means we cannot reject the claim that these two models are significantly different in performance.

We also report the coefficient of determination (R^2) for all our models: the best R^2 was 0.17 on the validation folds and 0.14 on the held-out test set. These low numbers are expected, given the relatively small size of this INTERSPEECH challenge data set and

TABLE 2 | Results of our best performing regression models predicting a subject's MMSE score (ranging from 0 to 30, with lower values indicating more severe dementia). The 'Baseline (Acoustic)' model is described in Luz et al. (2020). As in **Table 1**, the columns on the right show RMSE and R^2 on the held-out test set of 48 subjects (24 AD and 24 non-AD). The middle columns list RMSE and R^2 using 5-fold cross-validation on the training set of 108 subjects (54 AD and 54 non-AD).

Regressors	RMSE	R^2	RMSE (hold-out set)	R^2
Random forest (NLP)	6.43 ± 0.18	0.17	5.62	0.14
Gradient boosting (acoustic)	6.89 ± 0.17	0.06	6.67	−0.21
Random forest (NLP) + gradient boosting (acoustic)	6.66 ± 0.18	0.13	6.01	0.02
Majority vote (all models)	6.85 ± 0.16	0.10	6.12	−0.02
Baseline (acoustic)	7.30	–	6.14	–

the complexity of the condition. Interpreting this statistic in an absolute sense is problematic, especially as we did not find any other study using the same data set that reported this metric. We note that models based on language features achieved the best R^2 values, which further supports our claim that language features are very important for this task.

4 DISCUSSION

We investigated a variety of ML models, using language and/or acoustic features, to identify models that performed well at using speech information to distinguish AD from healthy subjects, and to estimate the severity of AD. Our results, of over 85% accuracy for classification and approximately 5.6 RMSE for regression, demonstrate the promise of using ML for detecting cognitive decline from speech. In our investigation, we explored multiple different combinations of features and ML algorithms; in the future, it would be interesting to delve deeper into the behavior of our best models, to determine the contribution of individual (or groups of) features to the model's ability to distinguish AD patients from healthy controls. Further, although we have currently used the full set of standard stopwords for removing noise in our language models, it may be worthwhile to see whether using a reduced set of stopwords (for example, preserving pronouns) might be more advantageous.

Our current best-performing models outperform recent results reported in the literature and provide evidence that, for discriminating between subjects with AD vs. healthy controls, features based on language (semantics, fluency and n-grams) are very useful. Compared to other top ranked results, our methods do not involve complex, computationally expensive algorithms. Instead, we used an ensemble approach with simple models to produce competitive results. Furthermore, a weighted majority vote of acoustic and language based models demonstrates competitive performance, implying that a combination of acoustic and language features also holds potential. Finally, comparing only acoustic models, we find that accuracy improves significantly compared to the baseline model (Luz et al., (2020)) for both the classification and regression tasks.

Our competitive performance, obtained using simple feature engineering along with classical machine learning algorithms, indicates that putting together an efficient machine learning pipeline from basic building blocks can achieve nearly state-of-the-art results for the learning tasks explored in this study. This result suggests that, for detecting AD from speech, it may be useful to explore traditional feature engineering and machine learning tools, especially in a limited data setting, as this will additionally provide for better interpretability and reproducibility compared to more complex deep learning based methods.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the DementiaBank corpus of the TalkBank repository [https://dementia.talkbank.org/].

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by TalkBank Code of Ethics Carnegie Mellon University [https://talkbank.org/share/ethics.html]. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of this study. ZS prepared the validation sets, and developed and tested the language based models. JS, MT, and SQ developed and tested the acoustic models. ZS, JS, MT, and SQ wrote sections of the manuscript. All authors contributed to manuscript editing and revision, and approved the submitted version.

FUNDING

RG and ZS gratefully acknowledge the funding from NSERC (Canada's Natural Science and Engineering Research Council) and Amii (Alberta Machine Intelligence Institute); and JS gratefully acknowledges the funding from IBM and MITACS. MT and ES gratefully acknowledge support from the AGE-WELL NCE; and MT gratefully acknowledges support from the Alberta Graduate Excellence Scholarship (AGES).

ACKNOWLEDGMENTS

The authors thank the ADReSS challenge organizing committee, including FH, SL, BM, and DF, for their efforts and support.

REFERENCES

Adlam, A.-L. R., Bozeat, S., Arnold, R., Watson, P., and Hodges, J. R. (2006). Semantic knowledge in mild cognitive impairment and mild alzheimer's disease. *Cortex* 42, 675–684. doi:10.1016/s0010-9452(08)70404-0

Alzheimer's Association (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's Demen.* 15, 321–387. doi:10.1016/j.jalz.2019.01.010

Arnáiz, E., and Almkvist, O. (2003). Neuropsychological features of mild cognitive impairment and preclinical alzheimer's disease. *Acta Neurol. Scand.* 107, 34–41. doi:10.1034/j.1600-0404.107.s179.7.x

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection.Preprint repository name [Preprint]. Available at: arXiv:2008.01551 (Accessed July 26, 2020).

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol.* 51, 585–594. doi:10.1001/archneur.1994.00540180063015

Braak, H., and Braak, E. (1991). Neuropathological stageing of alzheimer-related changes. *Acta Neuropathol.* 82, 239–259. doi:10.1007/BF00308809

Chakraborty, R., Pandharipande, M., Bhat, C., and Kopparapu, S. K. (2020). Identification of dementia using audio biomarkers. Preprint repository name [Preprint]. Available at: arXiv:2002.12788 (Accessed February 27, 2020).

Chien, Y.-W., Hong, S.-Y., Cheah, W.-T., Yao, L.-H., Chang, Y.-L., and Fu, L.-C. (2019). An automatic assessment system for alzheimer's disease based on speech using feature sequence generator and recurrent neural network. *Sci. Rep.* 9, 1–10. doi:10.1038/s41598-019-56020-x

De Jong, N. H., and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behav. Res. Methods* 41, 385–390. doi:10.3758/BRM.41.2.385

Dubois, B., Picard, G., and Sarazin, M. (2009). Early detection of alzheimer's disease: new diagnostic criteria. *Dialog. Clin. Neurosci.* 11, 135–139. doi:10.31887/DCNS.2009.11.2/bdubois

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM international conference on multimedia, Firenze, Italy, October 25–29, 2010 (New York, NY: AMC), 1459–1462.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *J. Alzheim. Dis.* 49, 407–422. doi:10.3233/JAD-150520

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *BDAE-3: Boston diagnostic Aphasia examination.* 3rd Edn. Philadelphia, PA: Lippincott Williams and Wilkins.

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., and Kálmán, J. (2010). Temporal parameters of spontaneous speech in alzheimer's disease. *Int. J. Speech Lang. Pathol.* 12, 29–34. doi:10.3109/17549500903137256

Jacobs, D. M., Sano, M., Dooneief, G., Marder, K., Bell, K. L., and Stern, Y. (1995). Neuropsychological detection and characterization of preclinical alzheimer's disease. *Neurology* 45, 957–962. doi:10.1212/wnl.45.5.957

Kempler, D. (1995). Language changes in dementia of the alzheimer type. *Demen. Commun.* 7, 98–114.

Kirshner, H. S. (2012). Primary progressive aphasia and alzheimer's disease: brief history, recent evidence. *Curr. Neurol. Neurosci. Rep.* 12, 709–714. doi:10.1007/s11910-012-0307-2

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition. Preprint repository name [Preprint]. Available at: arXiv:2009.04070 (Accessed September 09, 2020).

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the adress challenge. Preprint repository name [Preprint]. Available at: arXiv:2004.06833 (Accessed April 14, 2020).

MacWhinney, B. (2017). Tools for analyzing talk part 1: The chat transcription format. Available at: http//childes.psy.cmu.edu/manuals/CHAT.pdf (Accessed April 2014).

Miller, G. A. (1998). *WordNet: an electronic lexical database.* Cambridge, MA: MIT press, 449.

Mirzaei, S., El Yacoubi, M., Garcia-Salicetti, S., Boudy, J., Kahindo Senge Muvingi, C., Cristancho-Lacroix, V., et al. (2017). "Automatic speech analysis for early Alzheimer's disease diagnosis," in JETSAN 2017: 6e Journées d'Etudes sur la Télésanté, Bourges, France, May–June 31–01, 2017 (Bourges, France: JETSAN), 114–116.

Nestor, P. J., Scheltens, P., and Hodges, J. R. (2004). Advances in the early detection of alzheimer's disease. *Nat. Med.* 10, S34–S41. doi:10.1038/nrn1433

Nicholas, M., Obler, L. K., Albert, M. L., and Helm-Estabrooks, N. (1985). Empty speech in alzheimer's disease and fluent aphasia. *J. Speech Lang. Hear. Res.* 28, 405–410. doi:10.1044/jshr.2803.405

Ritchie, K., and Lovestone, S. (2002). The dementias. *Lancet* 360, 1759–1766. doi:10.1016/S0140-6736(02)11667-9

Sarawgi, U., Zulfikar, W., Soliman, N., and Maes, P. (2020). Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. Preprint repository name [Preprint]. Available at: arXiv:2009.00700 (Accessed August 30, 2020).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., et al. (2010). "The interspeech 2010 paralinguistic challenge," in Eleventh annual Conference of the international speech communication association, Makuhari, Chiba, Septmber 26–30, 2010 (Makuhari, Japan: ISCA), 3137.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in Proceedings INTERSPEECH 2013, 14th annual conference of the international speech communication association, Lyon, France, August 25-29, 2013 (France, EU: International Speech Communication Association (ISCA)), 3500.

Searle, T., Ibrahim, Z., and Dobson, R. (2020). Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech. Preprint repository name [Preprint]. Available at: arXiv:2006.07358 (Accessed June 12, 2020).

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Front. Aging Neurosci.* 7, 195. doi:10.3389/fnagi.2015.00195

Terry, R. D., Masliah, E., Salmon, D. P., Butters, N., DeTeresa, R., Hill, R., et al. (1991). Physical basis of cognitive alterations in alzheimer's disease: synapse loss is the major correlate of cognitive impairment. *Ann. Neurol.* 30, 572–580. doi:10.1002/ana.410300410

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* 15, 130–138. doi:10.2174/1567205014666171121114930

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., et al. (2013). "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, Barcelona, Spain, October, 2013 (New York, NY: ACM), 3–10.

Wankerl, S., Nöth, E., and Evert, S. (2017). "An n-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language," in Interspeech 2017, Stockholm, Sweden, August 20–24, 2017 (Stockholm, Sweden: ISCA), 3162–3166.

Weller, J., and Budson, A. (2018). Current understanding of alzheimer's disease diagnosis and treatment. *F1000Res.* 7, F1000. doi:10.12688/f1000research.14506.1

Conflict of Interest: The authors declare that this study received funding from the IBM Center for Advanced Studies. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

frontiers
in Psychology

BRIEF RESEARCH REPORT
published: 12 February 2021
doi: 10.3389/fpsyg.2020.623237

Towards Computer-Based Automated Screening of Dementia Through Spontaneous Speech

Karol Chlasta[1,2]* *and Krzysztof Wołk*[1]

[1] *Department of Computer Science, Polish-Japanese Academy of Information Technology, Warsaw, Poland,* [2] *Institute of Psychology, SWPS University of Social Sciences and Humanities, Warsaw, Poland*

Dementia, a prevalent disorder of the brain, has negative effects on individuals and society. This paper concerns using Spontaneous Speech (ADReSS) Challenge of Interspeech 2020 to classify Alzheimer's dementia. We used (1) VGGish, a deep, pretrained, Tensorflow model as an audio feature extractor, and Scikit-learn classifiers to detect signs of dementia in speech. Three classifiers (LinearSVM, Perceptron, 1NN) were 59.1% accurate, which was 3% above the best-performing baseline models trained on the acoustic features used in the challenge. We also proposed (2) DemCNN, a new PyTorch raw waveform-based convolutional neural network model that was 63.6% accurate, 7% more accurate then the best-performing baseline linear discriminant analysis model. We discovered that audio transfer learning with a pretrained VGGish feature extractor performs better than the baseline approach using automatically extracted acoustic features. Our DepCNN exhibits good generalization capabilities. Both methods presented in this paper offer progress toward new, innovative, and more effective computer-based screening of dementia through spontaneous speech.

Keywords: dementia detection, prosodic analysis, affective computing, transfer learning, convolutional neural network, machine learning, speech technology, mental health monitoring

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Fahim Salim,
University of Twente, Netherlands
Maria Koutsombogera,
Trinity College Dublin, Ireland

****Correspondence:***
Karol Chlasta
karol@chlasta.pl

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Psychology

Received: *29 October 2020*
Accepted: *30 December 2020*
Published: *12 February 2021*

Citation:
Chlasta K and Wołk K (2021) Towards Computer-Based Automated Screening of Dementia Through Spontaneous Speech.
Front. Psychol. 11:623237.
doi: 10.3389/fpsyg.2020.623237

1. INTRODUCTION

One of the most important social problems in developed countries is the constant rise of the percentage of the elderly population. A major health issue affecting this segment of population is the appearance Alzheimer's dementia (AD), affecting around 50 million people worldwide and expected to grow three times over the next 50 years (Baldas et al., 2010).

Dementia is estimated to be responsible for 11.2% of years lived with disability in people over 60 years of age, compared with 9.5% for stroke, 5.0% for cardiovascular disease, and 2.4% for cancer. In Europe, the prevalence of AD increases exponentially with age. The incidence also increases with age, although with a plateau in extreme old age (Todd and Passmore, 2009).

Comorbidity of several physical and mental health disorders was studied in relation to age and socioeconomic deprivation. The presence of mental health disorders increased as the number of physical morbidities increased, and was much greater in more deprived than in less deprived people. Physical-mental health comorbidity is very common, with depression and painful disorders as key comorbidities, and with dementia seen in a small reverse gradient (Barnett et al., 2012).

There is a significant relation between old-age depression and subsequent dementia in patients over the age of 50. This supports the hypothesis of old-age depression being a predictor, and possibly a causal factor of subsequent dementia (Buntinx et al., 1996).

Speech is a well-established early indicator of cognitive deficits including dementia (Bucks et al., 2000). Speech processing methods offer great potential to fully automatically screen for prototypic indicators in near real time, and they can be used as an additional information source when diagnosing Alzheimer's disease (Weiner et al., 2016).

Dementia was detected in speech with voice activity detection and speaker diarization followed by extraction of acoustic features. The unsupervised system achieved up to 0.645 unweighted average recall (UAR). Authors detected dementia using speech segments as short as 2.5 min, but achieved the best results using segments in the range between 10 and 15 min (Weiner et al., 2018).

Other AD detection approaches combined extraction of acoustic and linguistic features (Speech to Text and Human Transcriptionist), and applied a one-way ANOVA for feature selection. The reported binary classification accuracy on brief (less than 10 min) spontaneous speech samples reached 88%, with recall of 0.920 (Jarrold et al., 2014).

We target the classification task of AD Recognition through Spontaneous Speech (ADReSS Challenge 2020). The AD classification task consists of creating binary classification models to distinguish between AD and non-AD patient speech on the ADReSS dataset. The authors of that challenge prepared the dataset and provided five baseline, machine learning classification models, that used both acoustic and linguistic features for the detection of AD in spontaneous speech. Their acoustic approaches were based on emobase (Eyben et al., 2010), ComParE 2013 (Eyben et al., 2013), Multi-resolution Cochleagram features (MRCG) proposed by Chen et al. (2014), the Geneva minimalistic acoustic parameter set (eGeMAPS) by Eyben et al. (2015), and minimal feature set (Luz, 2017). The best baseline accuracy was achieved by linear discriminant analysis (LDA) model using ComParE features.

In this paper, we propose two methods for speech-based screening of AD. Our models perform significantly better than the ADReSS challenge baseline for classification task, as evaluated on the same, official ADReSS challenge dataset.

2. METHODS

2.1. Dataset

The dataset for the 2020 ADReSS challenge consists of speech recordings elicited for the Cookie Theft picture description task from the Boston Diagnostic Aphasia Exam (Goodglass et al., 2001). These data were balanced by the organizers in terms of age, gender, and the distribution of labels between the training and test partitions in order to minimize the risk of bias in the prediction tasks. The dataset from 78 non-AD subjects, and 78 AD subjects, was labeled for binary classification and regression tasks. The labels for the binary classification include Alzheimer's dementia and healthy control, whereas the labels for the regression task are Mini-Mental State Examination (MMSE) scores (Folstein et al., 1975), which provide a means for dementia diagnosis based on linguistic tests. For more details regarding the dataset, including the segmentation and voice activity detection algorithm, we refer the reader to the ADReSS challenge baseline paper (Luz et al., 2020).

2.2. VGGish Model and Scikit-Learn Classifiers

We extended the method of Pons Puig et al. (2018) and conducted two-step classification experiments to detect cognitive impairment due to AD (as shown in **Figure 1**). This consisted of a two-stage classification process, where a classifier was trained with features to predict whether a speech segment was uttered by a non-AD or AD patient, and majority vote (MV) classification, which assigned each subject an AD or non-AD label based on the majority labels classification.

2.2.1. Feature Extraction

We used VGGish (Hershey et al., 2017), a deep, pretrained Tensorflow (Abadi et al., 2016) model as a feature extractor. VGGish is an audio embedding produced by training a modified VGGNet model (Simonyan and Zisserman, 2014) to predict video tags from the Youtube-8M dataset (Abu-El-Haija et al., 2016). Principal component analysis (PCA) (Cao et al., 2003) was used for dimensionality reduction, with PCA set to 128. VGGish model converted audio input features into high-level 128-D embedding, which was fed as an input to a downstream classification model. The features were extracted from non-overlapping audio patches of 0.96 s, where each audio patch covered 64 mel bands and 96 frames of 10 ms each.

2.2.2. Classification Methods

We performed classification experiments using five different methods, namely support vector machines (SVM, with a radial basis function kernel and scaling gamm), linear support vector machines (LSVM), perceptron, multi-layer perceptron classifier (MLP, with 20 hidden layers, using a stochastic gradient descent solver, 600 iterations, learning rate of 0.001), and nearest neighbor (1NN, for KNN with $K = 1$ and cosine metric).

2.3. DemCNN—Custom Convolutional Neural Network

Current deep convolutional neural network (CNN) performs considerably better than the previous state-of-the-art (Krizhevsky et al., 2012). Transfer learning was often used in medical image analysis (Cheplygina et al., 2019). Applying transfer learning on a wide range of tasks nearly always gave better results (Kornblith et al., 2019). CNN-based methods have been successfully employed to medical imaging tasks and achieved human-level performance in classification tasks (Esteva et al., 2019). CNNs have proven very effective in image classification and show promise for audio (Hershey et al., 2017). We extend the audio classification work presented in Wołk, K., and Wołk (2019) and Chlasta et al. (2019).

2.3.1. Classification Method

We introduce DemCNN, a custom PyTorch (Paszke et al., 2019) CNN. We designed and implemented a custom sequential architecture consisting of six Conv1D layers using ReLU activation function, batch normalization and dropout, with

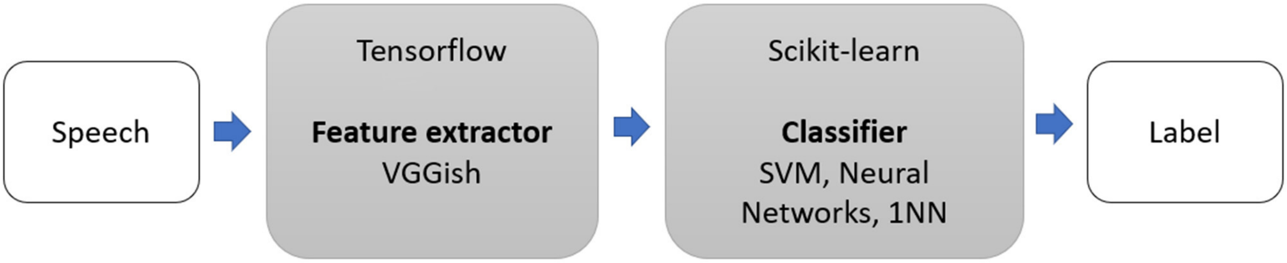


FIGURE 1 | Two-stage architecture: VGGish model and Scikit-learn classifiers.

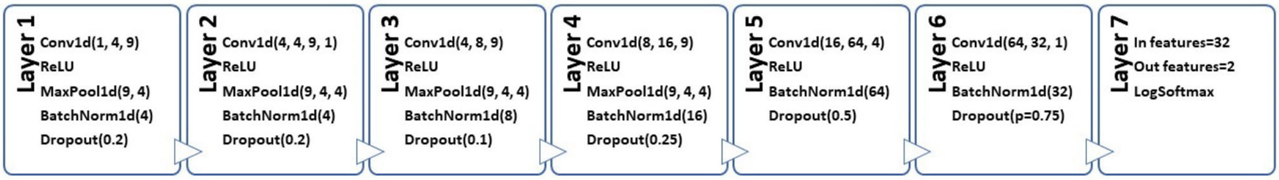


FIGURE 2 | Architecture diagram of DemCNN, a custom PyTorch convolutional neural network for speech classification.

the final (seventh) output layer being a dense layer. The output layer had 2 nodes (num_labels), which matched the number of possible classifications outputs. **Figure 2** presents a more detailed architecture diagram of our custom CNN for speech classification.

We unpacked a byte-string for each file into a 1D numpy array of numbers that could be analyzed by the CNN. Subsequently, the dataset was downsampled with a low-pass filter (with downsampling factors of 4, 4, 2).

We performed a two-step training of our CNN model using a cross-entropy loss function. We fine-tuned learning rate, the number of training cycles, and the number of training iterations per cycle. We set the first (training) batch size to 32, and the second (deployment) batch size to 2. The selection of the second learning rate for each step of our method was automated using a custom function operating on standard lr_finder. We trained the classifier for 2 or 4 epochs.

3. EXPERIMENTS AND RESULTS

All experiments were implemented in Python using Scikit-learn (Pedregosa et al., 2011), Tensorflow (Abadi et al., 2016), and PyTorch (Paszke et al., 2019) on the Google Colaboratory Platform (Bisong, 2019). The platform uses Jupyter Notebook standard that facilitates exchange of source code and reproducibility of results. The source code and accompanying results are available on GitHub.[1]

The ADReSS development data were split into train and test sets by randomly assigning 80% of the speakers to the train set and 20% to the test set. Results obtained for different classifier setups are summarized in **Table 1**.

Three models we developed using the first approach (VGGish + 128 PCA + linearSVM/perceptron/1NN) achieved 59% accuracy in our test set. Employing the same setup with SVN model, we achieved 55% accuracy. The best-performing baseline SVM models using MRCG features proposed by (Chen et al., 2014) and the ComParE 2013 features (Eyben et al., 2013) achieved lower accuracy of 53%. Interestingly, our 1NN model achieved better results than the best-performing baseline 1NN model using ComParE features (59% against 57%).

Our custom raw waveform DemCNN system achieved the best classification accuracy of 63.6%. The model classified 14 speakers correctly, eight incorrectly, and proved the most effective in distinguishing between AD and non-AD speech samples on the full wave enhanced ADReSS audio dataset. This result was 7% better then the best baseline classification accuracy on the ADReSS training set (Luz et al., 2020).

The final results for our custom audio DemCNN model were submitted to the 2020 ADReSS Challenge organizers after retraining the classifier on the full ADReSS training set, and predicting on the full ADReSS test set (see **Table 2** for results and the accompanying hyperparameters). Our model performed slightly better (1%) on the test partition than the best baseline LDA model trained on automatically extracted ComParE feature set (Eyben et al., 2013).

4. DISCUSSION

The main limitations of the AD field are poor standardization, limited comparability of results, and a degree of disconnect between study aims and clinical applications (de la Fuente Garcia

[1] Code: https://github.com/KarolChlasta/ADReSS-Challenge2020

TABLE 1 | Summary of classification results on AD Recognition through Spontaneous Speech (ADReSS) training set.

Model type	Precision	Recall	F1 score	Accuracy	Baseline accuracy
SVM	0.556	0.454	0.500	0.545	0.565 (SVM + Minimal)
LinearSVM	0.600	0.545	0.571	**0.591**	0.565 (SVM + Minimal)
Perceptron	0.600	0.545	0.571	**0.591**	0.565 (LDA + ComParE)
MLP	0.429	0.273	0.333	0.454	0.565 (LDA + ComParE)
1NN	0.600	0.545	0.571	**0.591**	0.574 (1NN + ComParE)
DemCNN	0.692	0.692	0.692	**0.636**	0.565 (LDA + ComParE)

Our approaches (VGGish + 128 Principal component analysis [PCA] and custom audio convolutional neural network [DemCNN]) vs. the best baseline accuracy on acoustic features. The bold values indicate best results achieved on ADReSS training dataset.

TABLE 2 | The results of Alzheimer's dementia (AD) classification task on AD Recognition through Spontaneous Speech (ADReSS) test set.

Approach	Class	Precision	Recall	F1 Score	Accuracy
DemCNN (Learning rate = 0.2;	Non-AD	0.528	0.792	0.633	0.542
Cycles = 4.4; Lengths = 8.8)	AD	0.528	0.792	0.389	0.542
DemCNN (Learning rate = 0.1;	Non-AD	0.625	0.625	0.625	**0.625**
Cycles = 2.2; Lengths = 8.8)	AD	0.625	0.625	0.625	**0.625**
Baseline acoustic features	Non-AD	0.670	0.500	0.570	0.620
(LDA + ComParE)	AD	0.600	0.750	0.670	0.620

Our approach (custom audio convolutional neural network [DemCNN]) vs. the best baseline on acoustic features (linear discriminant analysis [LDA] + CompParE). The bold values indicate best results achieved on ADReSS test dataset.

et al., 2020). Our two methods are attempting to close some of these gaps.

Data scarcity has hindered research into the relationship between speech and dementia. Recently, the community has turned to transfer learning (Yosinski et al., 2014), as a solution for a wide range of machine learning tasks for which labeled data are scarce. Selecting the right pretrained model as audio feature extractor allows to rapidly prototype competent speech classifiers.

In our first approach, we used a standard VGGish (Hershey et al., 2017), that is a popular deep audio embedding model trained on Youtube-8M video dataset (Abu-El-Haija et al., 2016). In our experiments to detect subtle changes in pathological speech, we confirmed that automatic extraction of acoustic features (Eyben et al., 2010) performs similarly to using a pretrained deep audio embedding model for feature extraction.

Similarly to us, Syed et al. (2020) also used VGGish deep acoustic embeddings in the ADReSS Challenge. They used other types of feature aggregation methods: (a) Fisher Vector encodings (FVs) and (b) Bag-of-Audio-Words (BoAW). Both achieved satisfactory results. Their VGGish and FVs model overperformed ours (59.1%) with 62.96% accuracy on the train partition, whereas their VGGish and BoAW model achieved even higher accuracy of 75%.

Our second method, the DemCNN model, for which we only performed a basic hyperparameter tuning, improved the classification results further. Moreover, the results achieved by DemCNN were similar in training and testing (63.6 vs. 62.5%), which is a good indicator of the lack of overfitting during the training process. This can be explained by a larger dropout defined in layers 5 and 6 of the network. An expected consequence of that is a good generalization capacity of our DemCNN model, which would positively impact the overall performance in clinical practice, when working with new data.

A similar approach to our DemCNN in the ADReSS Challenge was proposed by Cummins et al. (2020). Their raw segment based End-to-End CNN had four convolution layers, with the first convolution layer used to model voice source-related information or vocal tract information, such as formants. This approach achieved 71.3% accuracy on the training partition, but the reported result on the test partition was only 66.7%. Although this result is 4% better than our DemCNN, an expected consequence of a large difference between the results in training and test partitions is possibly a worse generalization capability of the network when working with new data.

An interesting opportunity for future research would be to use a combination of acoustic and linguistic features in detecting dementia. The latter approach, derived from automatic speech recognition (ASR) output, or from manual transcripts, had already been proven to detect dementia (Weiner et al., 2017), but relatively small gains were found when fusing acoustics and linguistics approaches (Cummins et al., 2020; Rohanian et al., 2020).

ADReSS Challenge 2020 helped to establish that although the linguistic systems outperforms the acoustic systems in AD (Cummins et al., 2020; Yuan et al., 2020), this result is unsurprising given that a human observer generated the transcripts manually, and they contain considerably fewer sources of noise than the audio recordings. As a result, such systems would be difficult to implement in clinical practice.

An option to overcome that would be to combine acoustic information with linguistics systems based on transcripts generated from ASR systems. This idea would introduce automation, but also increase the complexity, and dependency on errors rate for ASR in a given language.

It may also be useful for future work to gather a large dataset combining spontaneous speech samples for several pathologies (starting with depression and dementia, especially for old-age patients) to train an improved DepCNN to distinguish different types of disorders in pathological speech.

Finally, the DementiaBank's Pitt corpus (Jost and Grossberg, 1995) is large enough for considering experiments with other, custom, or off-the-shelf deep neural network architectures.

5. CONCLUSION

In this paper, we proposed and compared two acoustic-based systems: VGGish, a pretrained Tensorflow model as audio feature extractor and Scikit-learn classifiers with DemCNN, a custom raw waveform based CNN.

In the first approach, we selected the VGGish model as feature extractor and PCA for dimensionality reduction. This approach achieved the accuracy of 59.1%, 3% better than the best baseline accuracy achieved on the train partition with acoustic feature extraction for the respective classification algorithms.

In the second approach, we presented DemCNN, our custom PyTorch audio CNN to detect signs of dementia in spoken language. According to the experiments, the proposed architecture achieved promising performance and demonstrated the effectiveness of our method, as well as good generalization capabilities. DemCNN overperformed the best baseline accuracy of LDA model (ComParE feature set) by 7% on the ADReSS training set (accuracy of 63.6%), and 1% on the test ADReSS test set (accuracy of 62.5%). Our DemCNN and End-to-End Convolutional Neural Network (Cummins et al., 2020) produced the strongest performance of the acoustic systems on the ADReSS 2020 classification task, highlighting the benefits of self-learning features.

To conclude, we demonstrated a proof-of-concept, and applicability of (1) audio transfer learning for feature extraction, (2) DemCNN, a custom raw waveform based CNN in detecting dementia through spontaneous speech. We demonstrated that (1) audio transfer learning with a pretrained VGGish feature extractor performs better then the baseline approach (Luz et al., 2020) using automatically extracted acoustic features, and that these are relatively minor improvements. Our DemCNN method (2) overperforms our VGGish method (1) by 4% and the baseline on the test partition (Luz et al., 2020) by roughly 1%.

Both approaches presented are active attempts to close the gaps in standarization of automatic AD detection, and to improve the overall comparability of results to better embed computational speech technology into clinical practice. They offer simplicity, easy deployment, and they are language independent, which could result in a wide adoption and improved accessibility in a short space of time.

This contribution is especially important now, in the time of current COVID-19 pandemic, when the need for a remote digital health assessment tool is greater than ever for the elderly and other vulnerable populations.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://dementia.talkbank.org/.

AUTHOR CONTRIBUTIONS

KC was responsible for conceptualization, algorithmic development, data analysis, investigation, validation, and writing of original draft. KW supervises the entire work and contributes to idea conceptualization, algorithmic development, manuscript revision, and approval of the submission.

ACKNOWLEDGMENTS

We thank Michael Connolly for assistance with editorial reviews that improved the manuscript.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA), 265–283.

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., et al. (2016). Youtube-8m: a large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*. Available online at: https://research.google/pubs/pub45619/

Baldas, V., Lampiris, C., Capsalis, C., and Koutsouris, D. (2010). "Early diagnosis of Alzheimer's type dementia using continuous speech recognition," in *International Conference on Wireless Mobile Communication and Healthcare* (Berlin; Heidelberg: Springer), 105–110. doi: 10.1007/978-3-642-20865-2_14

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., and Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 380, 37–43. doi: 10.1016/S0140-6736(12)60240-2

Bisong, E. (2019). "Google colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (Berkeley, CA: Springer), 59–64. doi: 10.1007/978-1-4842-4470-8_7

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology* 14, 71–91. doi: 10.1080/026870300401603

Buntinx, F., Kester, A., Bergers, J., and Knottnerus, J. A. (1996). Is depression in elderly people followed by dementia? A retrospective cohort study based in general practice. *Age Ageing* 25, 231–233. doi: 10.1093/ageing/25.3.231

Cao, L., Chua, K. S., Chong, W., Lee, H., and Gu, Q. (2003). A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* 55, 321–336. doi: 10.1016/S0925-2312(03)00433-8

Chen, J., Wang, Y., and Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 1993–2002. doi: 10.1109/TASLP.2014.2359159

Cheplygina, V., de Bruijne, M., and Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* 54, 280–296. doi: 10.1016/j.media.2019.03.009

Chlasta, K., Wołk, K., and Krejtz, I. (2019). Automated speech-based screening of depression using deep convolutional neural networks. *Proc. Comput. Sci.* 164, 618–628. doi: 10.1016/j.procs.2019.12.228

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., et al. (2020). A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. *Proc. Interspeech* 2020, 2182–2186. doi: 10.21437/Interspeech.2020-2635

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J. Alzheimers Dis.* 78, 1547–1574. doi: 10.3233/JAD-200888

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2015). The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 190–202. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona), 835–838. doi: 10.1145/2502081.2502224

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze), 1459–1462. doi: 10.1145/1873951.1874246

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *BDAE-3: Boston Diagnostic Aphasia Examination, 3rd Edn.* Philadelphia, PA: Lippincott Williams and Wilkins.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., et al. (2017). "CNN architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA). doi: 10.1109/ICASSP.2017.7952132

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., et al. (2014). "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Baltimore, MD), 27–37. doi: 10.3115/v1/W14-3204

Jost, B. C., and Grossberg, G. T. (1995). The natural history of Alzheimer's disease: a brain bank study. *J. Am. Geriatr. Soc.* 43, 1248–1255. doi: 10.1111/j.1532-5415.1995.tb07401.x

Kornblith, S., Shlens, J., and Le, Q. V. (2019). "Do better imagenet models transfer better?" in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2661–2671. (Long Beach, CA: IEEE). 2661–2671. doi: 10.1109/CVPR.2019.00277

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Lake Tahoe, NV: Curran Associates, Inc.), 1097–1105.

Luz, S. (2017). "Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)* (Thessaloniki: IEEE), 45–46. doi: 10.1109/CBMS.2017.41

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). "Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge," in *Proceedings of INTERSPEECH 2020* (Shanghai). doi: 10.21437/Interspeech.2020-2571

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* 32: Annual Conference on Neural Information Processing Systems 2019, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC), 8024–8035. Available online at: https://proceedings.neurips.cc/paper/2019

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.org/papers/v12/pedregosa11a.html

Pons Puig, J., Nieto Caballero, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., and Serra, X. (2018). "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018* (Paris: International Society for Music Information Retrieval), 637–644.

Rohanian, M., Hough, J., and Purver, M. (2020). "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," in *Proc. Interspeech* (Shanghai), 2187–2191. doi: 10.21437/Interspeech.2020-2721

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. Available online at: https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/simonyan15.pdf

Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). "Automated screening for Alzheimer's dementia through spontaneous speech," in *Interspeech*, eds W. Hess and M. Cooke (Shanghai: International Speech Communication Association). 1–5. doi: 10.21437/Interspeech.2020-3158

Todd, S., and Passmore, P. (2009). Alzheimers disease, the importance of early detection. *Eur. Neurol. Rev.* 110, 18–21. doi: 10.17925/ENR.2008.03.02.18

Weiner, J., Angrick, M., Umesh, S., and Schultz, T. (2018). "Investigating the effect of audio duration on dementia detection using acoustic features," in *Interspeech*, eds W. Hess and M. Cooke (Hyderabad: International Speech Communication Association), 2324–2328. doi: 10.21437/Interspeech.2018-57

Weiner, J., Engelbart, M., and Schultz, T. (2017). "Manual and automatic transcriptions in dementia detection from speech," in *Interspeech*, eds W. Hess and M. Cooke (Stockholm: International Speech Communication Association), 3117–3121. doi: 10.21437/Interspeech.2017-112

Weiner, J., Herff, C., and Schultz, T. (2016). "Speech-based detection of Alzheimer's disease in conversational German," in *Interspeech*, eds W. Hess and M. Cooke (San Francisco, CA: International Speech Communication Association), 1938–1942. doi: 10.21437/Interspeech.2016-100

Wołk, K., and Wołk, A. (2019). Early and remote detection of possible heartbeat problems with convolutional neural networks and multipart interactive training. *IEEE Access* 7, 145921–145927. doi: 10.1109/ACCESS.2019.2919485

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, NeurIPS 2014*, eds Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger (Montréal, QC), 3320–3328. Available online at: https://proceedings.neurips.cc/paper/2014

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. *Proc. Interspeech* 2020, 2162–2166. doi: 10.21437/Interspeech.2020-2516

Conflict of Interest: The 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.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 24 March 2021
doi: 10.3389/fcomp.2021.624558

Recognition of Alzheimer's Dementia From the Transcriptions of Spontaneous Speech Using fastText and CNN Models

Amit Meghanani, C. S. Anoop and Angarai Ganesan Ramakrishnan*

MILE Laboratory, Department of Electrical Engineering, Indian Institute of Science, Bengaluru, India

Alzheimer's dementia (AD) is a type of neurodegenerative disease that is associated with a decline in memory. However, speech and language impairments are also common in Alzheimer's dementia patients. This work is an extension of our previous work, where we had used spontaneous speech for Alzheimer's dementia recognition employing log-Mel spectrogram and Mel-frequency cepstral coefficients (MFCC) as inputs to deep neural networks (DNN). In this work, we explore the transcriptions of spontaneous speech for dementia recognition and compare the results with several baseline results. We explore two models for dementia recognition: 1) fastText and 2) convolutional neural network (CNN) with a single convolutional layer, to capture the n-gram-based linguistic information from the input sentence. The fastText model uses a bag of bigrams and trigrams along with the input text to capture the local word orderings. In the CNN-based model, we try to capture different n-grams (we use $n = 2, 3, 4, 5$) present in the text by adapting the kernel sizes to n. In both fastText and CNN architectures, the word embeddings are initialized using pretrained GloVe vectors. We use bagging of 21 models in each of these architectures to arrive at the final model using which the performance on the test data is assessed. The best accuracies achieved with CNN and fastText models on the text data are 79.16 and 83.33%, respectively. The best root mean square errors (RMSE) on the prediction of mini-mental state examination (MMSE) score are 4.38 and 4.28 for CNN and fastText, respectively. The results suggest that the n-gram-based features are worth pursuing, for the task of AD detection. fastText models have competitive results when compared to several baseline methods. Also, fastText models are shallow in nature and have the advantage of being faster in training and evaluation, by several orders of magnitude, compared to deep models.

Keywords: fastText, convolutional neural network, Alzheimer's, dementia, mini-mental state examination

OPEN ACCESS

Edited by:
Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:
Diego R. Amancio,
University of São Paulo, Brazil
Anna Pribilova,
Slovak Academy of Sciences (SAS),
Slovakia

****Correspondence:***
C. S. Anoop
anoopcs@iisc.ac.in

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *31 October 2020*
Accepted: *25 January 2021*
Published: *24 March 2021*

Citation:
Meghanani A, Anoop CS and
Ramakrishnan AG (2021) Recognition
of Alzheimer's Dementia From the
Transcriptions of Spontaneous
Speech Using fastText and
CNN Models.
Front. Comput. Sci. 3:624558.
doi: 10.3389/fcomp.2021.624558

1 INTRODUCTION

Dementia is a syndrome characterized by the decline in cognition that is significant enough to interfere with one's independent, daily functioning. Alzheimer's disease contributes to around 60–70% of dementia cases. Toward the final stages of Alzheimer's dementia (AD), the patients lose control of their physical functions and depend on others for care. As there are no curative treatments for dementia, the early detection is critical to delay or slow down the onset or progression of the disease. The mini-mental state examination (MMSE) is a widely used test to screen for dementia and to estimate the severity and progression of cognitive impairment.

AD affects the temporal characteristics of spontaneous speech. Changes in the spoken language are evident even in mild AD patients. Subtle language impairments such as difficulties in word finding and comprehension, usage of incorrect words, ambiguous referents, loss of verbal fluency, speaking too much at inappropriate times, talking too loudly, repeating ideas, and digressing from the topic are common in the early stages of AD (Savundranayagam et al., 2005) and they turn extreme in the moderate and severe stages. Szatlóczki et al. (2015) show that AD can be detected with the help of a linguistic analysis more sensitively than with other cognitive examinations. Mueller et al. (2018b) analyzed the connected language samples obtained from simple picture description tasks and found that the speech fluency and the semantic content features declined faster in participants with early mild cognitive impairment. The language profile of AD patients is characterized by "empty speech," devoid of content words (Nicholas et al., 1985). They tend to use pronouns without proper noun references and indefinite terms like "this," "that," and "thing" more often (Mueller et al., 2018a). These results motivate us to believe that modeling the transcriptions of the narrative speech in the cookie-theft picture description task using n-gram language models can help in the detection of AD and prediction of MMSE score.

In this work we address the AD detection and MMSE score prediction problems using two natural language processing (NLP)–based models: 1) fastText and 2) convolutional neural network (CNN). These models have the advantage that they can be easily structured to capture the linguistic cues in the form of n-grams from the transcriptions of the picture description task, provided with the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset (Luz et al., 2020). CNNs, though originated in computer vision, have become popular for NLP tasks and have achieved great results in sentence classification (Kim, 2014), semantic parsing (tau Yih et al., 2014), search query retrieval (Shen et al., 2014), and other traditional NLP tasks (Collober et al., 2011). Our convolutional neural network model draws inspiration from the work on sentence classification using CNNs (Kim, 2014). The fastText (Joulin et al., 2017) is a simple and efficient model for text classification (e.g., tag prediction and sentiment analysis). The fundamental idea in the fastText classifier is to calculate the n-grams of an input sentence and append them to the end of the sentence. Our choice of fastText model is also motivated by its ability to often outperform deep learning classifiers in terms of accuracy and training/evaluation times (Joulin et al., 2017).

The rest of the paper is organized as follows. **Section 2** discusses the ADReSS dataset in detail. **Section 3** discusses the baseline results in AD detection. **Section 4** discusses our proposed NLP-based models followed by the listing of results in **Section 5**. Our results and conclusions are discussed in **Section 6**.

2 ADRESS DATASET

The ADReSS dataset (Luz et al., 2020) is designed to provide Alzheimer's research community with a standard platform for AD detection and MMSE score prediction. The dataset is acoustically preprocessed and balanced in terms of age and gender. It consists of audio recordings and transcriptions [in CHAT format (Macwhinney, 2009)] of the cookie-theft picture description task, elicited from subjects in the age group of 50–80 years. The training set consists of data from 108 subjects, 54 each from AD and non-AD classes. The test set has data from 48 subjects, again balanced with respect to AD and non-AD classes. More information on the ADReSS dataset can be found in the ADReSS challenge baseline paper (Luz et al., 2020).

3 REVIEW OF BASELINE METHODS

This section provides a brief overview of the various approaches for AD detection and MMSE score prediction on ADReSS dataset. These approaches can be broadly classified into three types based on the type of the features used in the problem: 1) acoustic feature, 2) linguistic feature, and 3) a fusion of acoustic and linguistic features. The performance of different approaches on the AD detection and MMSE score prediction tasks are compared using the accuracy and root mean square error (RMSE) measures computed on the ADReSS test set.

$$\text{Accuracy} = \frac{TN + TP}{N} \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \tag{2}$$

where N is the total number of subjects involved in the study, TP the number of true positives, and TN the number of true negatives. $\hat{y}_i$ and y_i are the estimated and target MMSE scores for i^{th} test sample. The results of different approaches on the ADReSS dataset are summarized in **Table 1**.

3.1 Acoustic Feature-Based Methods

Luz et al. (2020) explore several acoustic features like extended Geneva minimalistic acoustic parameter set (eGeMAPS) (Eyben et al., 2016), emobase, ComParE-2013 (Eyben et al., 2013), and multiresolution cochleagram (MRCG) (Chen et al., 2014), feeding the traditional machine learning algorithms like linear discriminant analysis, decision trees, nearest neighbor, random forests, and support vector machines. In our previous work (Meghanani et al., 2021), we have used CNN/ResNet + long short-term memory (LSTM) networks and pyramidal bidirectional LSTM + CNN networks trained on log-Mel spectrogram and Mel-frequency cepstral coefficient (MFCC) features extracted from the spontaneous speech. Pompili et al. (2020) exploit the pretrained models to produce i-vector- and x-vector-based acoustic feature embeddings. They evaluate x-vector, i-vector, and statistical speech-based functional features. Rhythmic features are proposed in Campbell et al. (2020), as lower speaking fluency is a common pattern in patients with AD. Koo et al. (2020) use VGGish (Hershey et al., 2017) trained with Audio Set (Gemmeke et al., 2017) for audio classification. They have proposed a modified version of convolutional recurrent neural network (CRNN), where an

TABLE 1 | Baseline methods on ADReSS test set.

Model	Accuracy (%)	RMSE
Searle et al. (2020), DistilBERT	81.25	4.58
Searle et al. (2020), SVM + CRF	81.25	5.22
Pompili et al. (2020), x-vectors SRE	54.17	—
Pompili et al. (2020), sentence embedding	72.92	—
Pompili et al. (2020), fusion of system	81.25	—
Luz et al. (2020), linguistic	75.00	5.20
Sarawgi et al. (2020b), ensemble	83.33	4.60
Koo et al. (2020), VGGish	72.92	5.07
Koo et al. (2020), Transformer-XL	81.25	4.01
Koo et al. (2020), VGGish + GloVe	77.08	4.33
Koo et al. (2020), VGGish + transformer-XL	75.00	3.74
Koo et al. (2020), ensembled output	81.25	3.77
Campbell et al. (2020), fusion II	75.00	—
Campbell et al. (2020), fusion I	72.92	—
Campbell et al. (2020), RNN model	75.00	—
Campbell et al. (2020), fluency	60.42	—
Campbell et al. (2020), x-vector	54.17	—
Sarawgi et al. (2020a), UA ensemble	—	4.35
Sarawgi et al. (2020a), UA ensemble (weighted)	—	3.93
Pappagari et al. (2020), acoustic and transcript	75.00	5.37
Rohanian et al. (2020), LSTM (Lexical + Dis)	72.92	4.88
Rohanian et al. (2020), LSTM with gating (Acoustic + Lexical)	77.08	4.57
Rohanian et al. (2020), LSTM with gating (Acoustic + Lexical + Dis)	79.17	4.54
Yuan et al. (2020), ERNIE3p	89.58	—
Syed et al. (2020)	85.42	4.30
Edwards et al. (2020), phonemes and audio	79.17	—
Meghanani et al. (2021), CNN-LSTM with MFCC	64..58	6.24
Meghanani et al. (2021), pBLSTM-CNN with log-Mel	52.08	5.90
Meghanani et al. (2021), ResNet-LSTM with log-Mel	62.50	5.98

attention layer is the forefront layer of the network, and fully connected layers follow the recurrent layer.

3.2 Linguistic Feature-Based Methods

Recently, there have been multiple attempts on the AD detection problem based on text-based features and models. Searle et al. (2020) use traditional machine learning techniques like support vector machines (SVMs), gradient boosting decision trees (GBDT), and conditional random fields (CRFs). They also try deep learning transformer-based models, specifically, bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT/DistilRoBERTa (Sanh et al., 2019). Pompili et al. (2020) encode each word of the clean transcriptions into 768-dimensional context embedding vector using a frozen English BERT model pretrained with 12 layers. Three different neural models are trained on top of contextual word embeddings: 1) global maximum pooling, 2) bidirectional long short-term memory (BLSTM)–based recurrent neural networks (RNN) provided with an attention module, and 3) the second model augmented with part-of-speech (POS) embeddings. In the work of Campbell et al. (2020), authors have used the manual transcripts to extract linguistic information (interventions, vocabulary richness, frequency of verbs, nouns, POS-tagging, etc.) for creating the input features of the classifier. They use another sequential deep learning-based classifier, which directly classifies the sequence of Gobal Vectors (GloVe)–based word embeddings. Koo et al. (2020) use transformer-based language models (Vaswani et al., 2017), generative pretraining (GPT) (Radford et al., 2018), RoBERTa (Liu et al., 2019), and transformer-XL (Dai et al., 2020) to get textual features and perform classification and regression tasks using a modified convolutional recurrent neural network-based structure.

Graph-based representation of word features (Tomás and Radev, 2012; Cong and Liu, 2014), which have shown promise in classifying texts (De Arruda et al., 2016), is also employed for detection of mild cognitive impairments. Santos et al. (2017) model transcripts as complex networks and enrich them with word embedding to better represent short texts produced in neuropsychological assessments. They use metrics of topological properties of complex networks in a machine learning classification approach to distinguish between healthy subjects and patients with mild cognitive impairments. Such graph-based techniques have also been used in the word sense disambiguation (WSD) tasks to identify the meaning of words in a given context for specific words conveying multiple meanings. Corra et al. (2018) suggest that a bipartite network model with local features employed to characterize the context can be useful in improving the semantic characterization of written texts without the use of deep linguistic information.

3.3 Bimodal Methods

Methods with bimodal input features (both acoustic and linguistic) are also used for AD recognition in various studies

(Sarawgi et al., 2020a; Sarawgi et al., 2020b; Campbell et al., 2020; Koo et al., 2020; Pompili et al., 2020; Rohanian et al., 2020). However, in this work, we restrict ourselves to the NLP-based approaches.

4 PROPOSED NLP-BASED METHODS

4.1 Data Preparation

In this work, we explore the linguistic features for AD detection and hence only the textual transcripts in the ADReSS dataset are used. The transcripts contain the conversational content between the participant and the investigator. This includes pauses in speech, laughter, and discourse markers such as "um" and "uh." Each transcript is considered as a single data point with their corresponding AD label and MMSE score. We create two transcription level datasets after preprocessing the transcripts as in Searle et al. (2020)—1) PAR: containing the utterances of participant alone, 2) PAR + INV: containing utterances from both the participant and the investigator. In addition to the preprocessing performed in Searle et al. (2020), we keep PAR and INV tags as well in the data (which defines whether the utterance is spoken by the participant or the investigator).

4.2 Convolutional Neural Network Model

Language impairments like difficulties in lexical retrieval, loss of verbal fluency, and breakdown in comprehension of higher order written and spoken languages are common in AD patients. Hence the linguistic information, like the n-grams present in the input sentence, may provide good cues for AD detection. Any $n \times d$ CNN filter, where n is the number of sequential words looked over by the filter and d is the dimension of word embedding, can be viewed as a feature detector looking for a specific n-gram in the input that can capture the language impairments associated with AD.

We describe the details of the CNN model from the work (Kim, 2014) as follows. Let $z_i \in R^d$ be a d-dimensional word vector corresponding to the ith word in the sentence. A sentence of length L is represented as $\{z_1, z_2, \ldots, z_L\}$. Let $z_{i:i+j}$ represent the concatenation of the words $z_i, z_{i+1}, \ldots, z_{i+j}$. A convolution operation involves a filter $w \in R^{\mathrm{nd}}$, which is applied to a window of n words to produce a new feature as shown in **Eq. 3**, where s_i is generated from a window of words $z_{i:i+n-1}$ by

$$s_i = f\,(w \cdot z_{i:i+n-1} + b). \tag{3}$$

In **Eq. 3**, f is a nonlinear function and b is the bias term. A feature map $\mathcal{E}$ is obtained by applying the filter to all possible windows of words in the sentence $[z_{1:n}, z_{2:n+1}, \ldots, z_{L-n+1:L}]$.

$$\mathcal{E} = [s_1, s_2, \ldots, s_{L-n+1}]. \tag{4}$$

A max-pool over time (Collober et al., 2011) is performed over the feature map to get $s_{\max} = \max\mathcal{E}$ as the feature corresponding to that filter. This corresponds to the n-gram that is "most relevant" in the AD recognition task. The weights of the filters, which in turn determine the "most relevant" feature, are learnt using backpropagation. CNNs are trained with just one layer of convolution. Variable length sentences are automatically handled by the pooling scheme. We use pretrained 100-dimensional GloVe word vectors (Pennington et al., 2014) for word embedding. Multiple kernels of sizes 2×100, 3×100, 4×100, and 5×100 are employed to have a look at the bigrams, trigrams, 4-grams, and 5-grams within the text. We use 100 filters each with heights 2, 3, 4, and 5. Multiple configurations with filter sizes [2,3,4], [3,4,5], and [2,3,4,5] are applied which are referred to as CNN-bi+tri+4 gram, CNN-tri+4+5 gram, and CNN-bi+tri+4+5 gram in our tables. The outputs of the filter are concatenated together to form a single vector. Dropout with probability $p = 0.5$ is applied on the concatenated filter output and the results are passed through a linear layer for the final prediction task. The linear layer weights up the evidence from each of these n-grams and make a final decision. **Figure 1** shows the basic CNN operation over an example sentence.

4.2.1 Training Details

For the classification task, training is performed for 100 epochs with a batch size of 16. Adam optimizer is used with a learning rate of 0.001. Model with the lowest validation loss is saved and

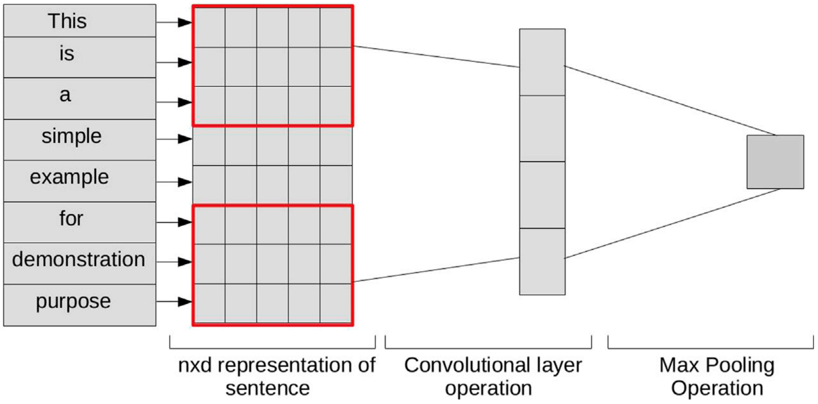


FIGURE 1 | Demonstration of CNN over text for an example sentence.

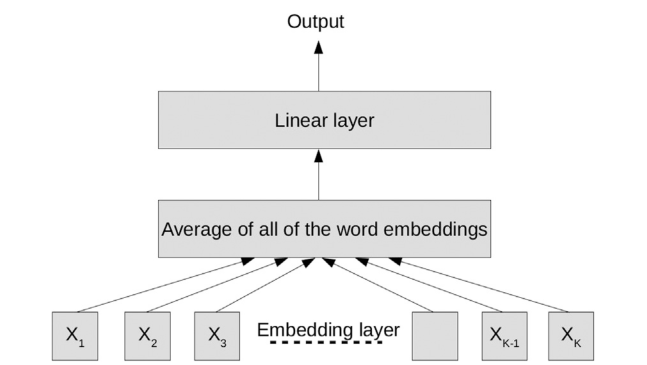


FIGURE 2 | fastText model (Joulin et al., 2017) with appended n-gram features (X_1,X_2,X_3,…,X_{K-1},X_K) as input.

used for prediction. Since AD classification is a two-class problem, binary cross-entropy with logits loss is used as the loss function. For the MMSE score prediction task, the output layer is a fully connected layer with linear activation function. In the regression task the network is trained for 1,500 epochs with the objective to minimize the mean squared error.

We use bootstrap aggregation of models known as bagging (Breiman, 1996) to predict the final labels/MMSE scores for test samples. Bootstrap aggregation is an ensemble technique to improve the stability and accuracy of machine learning models. It combines the prediction from multiple models. It also reduces variance and helps to avoid overfitting. We fit 21 models and the outputs are combined by a majority voting scheme for final classification. In the regression task, the outputs of these bootstrap models are averaged to arrive at the final MMSE score.

4.3 fastText

fastText-based classifiers calculate the n-grams of an input sentence explicitly and append them to the end of the sentence. In this work, we use bigrams and trigrams. We conducted the experiments with 4-grams as well, but the results did not show any improvement over the use of trigrams. This bag of bigrams and trigrams acts as additional features to capture some information about the local word order.

Figure 2 shows the architecture of fastText model. The fastText model has two layers, an embedding layer and a linear layer. The embedding layer calculates the word embedding (100-dimensional) for each word. The average of all these word embeddings is calculated and fed through the linear layer for final prediction as described in **Figure 2**. fastText models are faster for training and evaluation by many orders of magnitude, compared to the "deep" models. As mentioned in the work (Joulin et al., 2017), fastText can be trained on more than one billion words in less than 10 min using a standard multicore CPU and classify half a million sentences among 312 K classes in less than a minute.

4.3.1 Training Details

All training details are the same as mentioned in **Section 4.2.1**. The only difference is that dropout is not used in this model. Here also we use 21 bootstrapping models and the outputs are combined as described in **Section 4.2.1**.

TABLE 2 | Average 5-fold cross-validation results for AD classification and RMSE values.

Dataset	Model	Accuracy	RMSE
PAR	CNN, bi+tri+4 gram	73.91	4.55
PAR	CNN, tri+4+5 gram	77.54	4.41
PAR	CNN, bi+tri+4+5 gram	76.54	4.65
PAR	fastText, bigram	80.54	5.43
PAR	fastText, bi + trigram	82.36	5.40
PAR + INV	CNN, bi+tri+4 gram	80.18	4.63
PAR + INV	CNN, tri+4+5 gram	81.27	4.53
PAR + INV	CNN, bi+tri+4+5 gram	80.36	4.38
PAR + INV	fastText, bigram	86.09	4.66
PAR + INV	fastText, bi + trigram	85.90	4.81

5 RESULTS

We have performed 5-fold cross-validation, to estimate the generalization error. One of the folds has 20 validation samples and the remaining four have 22 validation samples. The results of cross-validation on CNN and fastText models trained on PAR and PAR + INV sets are listed in **Table 2**. The best performing model for classification during the cross-validation was fastText with bigrams on the PAR + INV set, which yields an average cross-validation accuracy of 86.09%. Among the CNN models, tri+4+5 grams give the best accuracy in both PAR (77.54%) and INV + PAR (81.27%) sets. As far as accuracy is concerned, both the CNN and fastText models seem to benefit from the inclusion of utterances from the investigator. For the prediction of MMSE score, CNN with bi+tri+4+5 grams (RMSE of 4.38) was the best. The fastText models seem to get a clear advantage in RMSE with the addition of the utterances from the investigator. However such a large difference in RMSE is not observable between the CNN models using PAR and INV + PAR sets. The cross-validation results confirmed our belief that the n-grams from the transcriptions of the picture description task could be useful in the detection of AD.

Table 3 lists the classification accuracy and RMSE in the prediction of MMSE score on the test set of the ADReSS corpus. The table also lists the precision, recall, and F_1score for each class. They are computed as precision $\pi = (TP/(TP + FP))$, recall $\rho = (TP/TP + FN)$, and F_1score = $(2\pi\rho/(\pi + \rho))$, where TP, FP, TN, and FN are the number of true positives, false positives, true negatives, and false negatives, respectively. The listed results are obtained after bootstrapping with 21 samples. The best classification accuracy is 83.33% which is achieved using fastText model with appended bigrams and trigrams. The accuracies are similar in both PAR and PAR + INV sets using the fastText model. The maximum accuracy obtained with CNN models is 79.16%, which is achieved on the INV + PAR set using bi+tri+4 grams or tri+4+5 grams. In the detection task, the CNN models seem to benefit from the addition of utterances from the investigator. Also the accuracies seem to degrade when bigrams,

TABLE 3 | Results on ADReSS test set. The bold values represent the best results obtained by our models.

Dataset	Model	Class	Precision	Recall	F1 score	Accuracy (%)	RMSE
PAR	CNN, bi+tri+4 gram	Non-AD	0.74	0.71	0.72	72.91	4.38
		AD	0.72	0.75	0.73		
PAR	CNN, tri+4+5 gram	Non-AD	0.76	0.67	0.71	72.91	4.46
		AD	0.70	0.79	0.75		
PAR	CNN, bi+tri+4+5 gram	Non-AD	0.71	0.71	0.71	70.83	4.42
		AD	0.71	0.71	0.71		
PAR	fastText, bigram	Non-AD	0.78	0.88	0.82	81.25	4.51
		AD	0.86	0.75	0.80		
PAR	fastText, bi + trigram	Non-AD	0.81	0.88	0.84	**83.33**	4.87
		AD	0.86	0.79	0.83		
PAR + INV	CNN, bi+tri+4 gram	Non-AD	0.77	0.83	0.80	79.16	4.48
		AD	0.82	0.75	0.78		
PAR + INV	CNN, tri+4+5 gram	Non-AD	0.77	0.83	0.80	79.16	4.47
		AD	0.82	0.75	0.78		
PAR + INV	CNN, bi+tri+4+5 gram	Non-AD	0.74	0.71	0.72	72.91	4.44
		AD	0.72	0.75	0.73		
PAR + INV	fastText, bigram	Non-AD	0.78	0.88	0.82	81.25	**4.28**
		AD	0.86	0.75	0.80		
PAR + INV	fastText, bi + trigram	Non-AD	0.79	0.92	0.85	**83.33**	4.47
		AD	0.90	0.75	0.82		

trigrams, 4-grams, and 5-grams are considered together. This behavior is consistent across the PAR and PAR + INV sets. The best RMSE in the prediction of MMSE score is 4.28 which is obtained on the PAR + INV set using fastText model employing only bigrams. In the regression task using fastText, the use of bigrams achieves slightly better RMSE compared to the use of both bigrams and trigrams. Also the fastText models seem to benefit from the use of utterances from the investigator. In contrast, CNN models do not seem to get any specific advantage with the inclusion of investigator's utterances. The performance of the CNN models remains almost the same across the use of bi+tri+4, tri+4+5, and bi+tri+4+5 grams.

6 DISCUSSION AND CONCLUSION

In this work, we explore two models, CNN with a single convolution layer and fastText, to address the problem of AD classification and prediction of MMSE score from the transcriptions of the picture description task. The choice of these models was based on our initial belief that modeling the transcriptions of the narrative speech in the picture description task using n-grams could give some indication on the status of AD. The chosen models are also shallow. The number of parameters is much less than the usual deep learning architectures and hence they can be trained and evaluated quite fast. Yet, the performance of these models is competitive with the baseline results reported with complex models (refer to **Table 1**). The results suggest that the n-gram-based features are worth pursuing, for the task of AD detection.

Among the considered models, fastText model with bigrams and trigrams appended to the input achieves the best classification accuracy (83.33%). In the regression task, the best results (RMSE of 4.28) are achieved using fastText model with only the bigrams appended to the input. The fastText models have a clear edge over CNN in the classification task. Empirical evidence suggests that fastText models benefit from the inclusion of utterances from the investigator in the regression task, though they do not make much difference in the classification task. The CNN models on the other hand perform better on the PAR + INV sets in the classification task. In the regression task, their performance is similar across the PAR and PAR + INV sets. Bigrams have an edge over bi + tri grams in fastText, when used for prediction of MMSE score. However, the performance of the CNN models remains almost the same across the use of bi+tri+4, tri+4+5, and bi+tri+4+5 grams, in the regression task.

DATA AVAILABILITY STATEMENT

The data analyzed in this study are subject to the following licenses/restrictions: In order to gain access to the ADReSS data, you will need to become a member of DementiaBank (free of charge) by contacting Brian MacWhinney on macw@cmu.edu. You should include your contact information and affiliation, as well as a general statement on how you plan to use the data, with specific mention to the ADReSS challenge. If you are a student, please ask your supervisor to join as a member as well. This membership will give you full access to the DementiaBank database, where the ADReSS dataset will be available and clearly identified. For further information, visit DementiaBank. Requests to access these datasets should be directed to Brian MacWhinney, macw@cmu.edu.

AUTHOR CONTRIBUTIONS

AM, AS, and AR contributed to the conception and design of the study. AM and AS wrote the first draft of the manuscript. AR reviewed the first draft and suggested improvements. AM and AS wrote sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

REFERENCES

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi:10.1007/BF00058655

Campbell, E. L., Docío-Fernández, L., Raboso, J. J., and García-Mateo, C. (2020). Alzheimer's dementia detection from audio and text modalities. arXiv preprint arXiv:2008.04617

Chen, J., Wang, Y., and Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 1993–2002. doi:10.1109/TASLP.2014.2359159

Collober, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Machine Learn. Res.* 12, 2493–2537. doi:10.5555/1953048.2078186

Cong, J., and Liu, H. (2014). Approaching human language with complex networks. *Phys. Life Rev.* 11, 598–618. doi:10.1016/j.plrev.2014.04.004

Corra, E. A., Lopes, A. A., and Amancio, D. R. (2018). Word sense disambiguation. *Inf. Sci.* 442, 103–113. doi:10.1016/j.ins.2018.02.047

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). "Transformer-XL: attentive language models beyond a fixed-length context," in Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy, July 2019, 2978–2988. doi:10.18653/v1/P19-1285

De Arruda, H., Costa, L., and Amancio, D. (2016). Using complex networks for text classification: discriminating informative and imaginative documents. *EPL* 113, 28007. doi:10.1209/0295-5075/113/28007

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, Minneapolis, MN, June 2–7, 2019, Vol. 1, 4171–4186. doi:10.18653/v1/N19-1423

Edwards, E., Dognin, C., Bollepalli, B., and Singh, M. (2020). "Multiscale system for Alzheimer's dementia recognition through spontaneous speech," in Proceedings of interspeech 2020, Shanghai, China, October 2020, 2197–2201. doi:10.21437/Interspeech.2020-2781

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affective Comput.* 7, 190–202. doi:10.1109/taffc.2015.2457417

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in Proceedings the 2013 ACM multimedia conference, Barcelona, Spain, October, 2013, 835–838. doi:10.1145/2502081.2502224

Gemmeke, J., Ellis, D., Freedman, D., Jansen, A., Lawrence, W., Moore, R., et al. (2017). "Audio set: an ontology and human-labeled dataset for audio events," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), New Orleans, LA, March 5–9, 2017, 776–780. doi:10.1109/ICASSP.2017.7952261

Hershey, S., Chaudhuri, S., Ellis, D., Gemmeke, J., Jansen, A., Moore, R. C., et al. (2017). "CNN architectures for large-scale audio classification," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), New Orleans, LA, March 5–9, 2017, 131–(135.)

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). "Bag of tricks for efficient text classification," in Proceedings of the 15th conference of the european chapter of the association for computational linguistics, Valencia, Spain, April 3–7, 2017, Vol. 2, 427–(431.)

Kim, Y. (2014). "Convolutional neural networks for sentence classification," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, October 25–29, 2014, 1746–1751. doi:10.3115/v1/D14-1181

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). "Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition," in Proceedings of interspeech 2020, Shanghai, China, October 2020, 2217–2221. doi:10.21437/Interspeech.2020-3153

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. ArXiv abs/1907.11692

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). "Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge," in Proceedings of interspeech 2020, Shanghai, China, October 2020, 2172–2176. doi:10.21437/Interspeech.2020-2571

Macwhinney, B. (2009). "The CHILDES project part 1," in *The CHAT transcription format*. doi:10.1184/R1/6618440.v1

Meghanani, A., Anoop, C. S., and Ramakrishnan, A. G. (2021). "An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech," in The 8th IEEE spoken language technology workshop (SLT), Shenzhen, China, January 19-22, 2021

Mueller, K. D., Hermann, B., Mecollarib, J., and Turkstra, L. S. (2018a). Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. *J. Clin. Exp. Neuropsychol.* 40, 917–939. doi:10.1080/13803395.2018.1446513

Mueller, K. D., Koscik, R. L., Hermann, B., Johnson, S. C., and Turkstra, L. S. (2018b). Declines in connected language are associated with very early mild cognitive impairment: results from the Wisconsin registry for Alzheimer's prevention. *Front. Aging Neurosci.* 9, 437. doi:10.3389/fnagi.2017.00437

Nicholas, M., Obler, L. K., Albert, M., and Helm-Estabrooks, N. (1985). Empty speech in Alzheimer's disease and fluent aphasia. *J. Speech Hear. Res.* 28, 405–410. doi:10.1044/jshr.2803.405

Pappagari, R., Cho, J., Moro-Velázquez, L., and Dehak, N. (2020). "Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity," in Proceedings of interspeech 2020, Shanghai, China, October 2020, 2177–2181. doi:10.21437/Interspeech.2020-2587

Pennington, J., Socher, R., and Manning, C. (2014). Glove: global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, October 25–29, 2014, 1532–1543. doi:10.3115/v1/d14-1162

Pompili, A., Rolland, T., and Abad, A. (2020). "The INESC-ID multi-modal system for the ADReSS 2020 challenge," in Proceedings of interspeech 2020, Shanghai, China, October 2020, 2202–2206. doi:10.21437/Interspeech.2020-2833

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. Available at: https://www.cs.ubc.ca/amuham01/LING530/papers/radford2018improving.pdf. doi:10.1017/9781108552202

Rohanian, M., Hough, J., and Purver, M. (2020). "Multi-Modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," in Proceedings of interspeech 2020, Shanghai, China, October 2020, 2187–2191. doi:10.21437/Interspeech.2020-2721

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv abs/1910.01108

Santos, L., Corrêa Júnior, E. A., Oliveira, O., Jr., Amancio, D., Mansur, L., and Aluísio, S. (2017). "Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts," in Proceedings the 55th annual meet: the association for computational linguistics, Vancouver, BC, July 30–August 4, 2017, Vol. 1, 1284–1296. doi:10.18653/v1/P17-1118

Sarawgi, U., Zulfikar, W., Khincha, R., and Maes, P. (2020a). Uncertainty-aware multi-modal ensembling for severity prediction of Alzheimer's dementia. ArXiv abs/2010.01440. doi:10.21437/interspeech.2020-3137

Sarawgi, U., Zulfikar, W., Soliman, N., and Maes, P. (2020b). Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity. arXiv preprint arXiv:2009.00700. doi:10.21437/interspeech.2020-3137

Savundranayagam, M., Hummert, M. L., and Montgomery, R. (2005). Investigating the effects of communication problems on caregiver burden. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 60 (1), S48–S55. doi:10.1093/geronb/60.1.s48

Searle, T., Ibrahim, Z., and Dobson, R. (2020). "Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech," in Proceedings of interspeech 2020, Shanghai, China, October 2020, 2192–2196. doi:10.21437/Interspeech.2020-2729

Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. (2014). "Learning semantic representations using convolutional neural networks for web search," in WWW 2014, Seoul, South Korea, April 7–11, 2014, 373–374. doi:10.1145/2567948.2577348

Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). "Automated screening for Alzheimer's dementia through spontaneous speech," Proceedings of interspeech 2020, Shanghai, China, October 2020, 2222–2226. doi:10.21437/Interspeech.2020-3158

Szatlóczki, G., Hoffmann, I., Vincze, V., Kálmán, J., and Pákáski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Front. Aging Neurosci.* 7, 110. doi:10.3389/fnagi.2015.00195

tau Yih, W., He, X., and Meek, C. (2014). "Semantic parsing for single-relation question answering," in Proceedings of the 52nd annual meeting of the association for computational linguistics, Baltimore, MA, June 2014, Vol. 2, 643–648. doi:10.3115/v1/P14-2105

Tomás, D. R. M., and Radev, D. (2012). Graph-based natural language processing and information retrieval. *Machine Translation* 26, 277–280. doi:10.1007/s10590-011-9122-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., et al. (2017). "Attention is all you need," Proceedings of the 31st international conference on neural information processing systems, Long Beach, CA, December 2017, 5999–(6009.)

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease," in Proceedings of interspeech 2020, Shanghai, China, October 2020, 2162–2166. doi:10.21437/Interspeech.2020-2516

Conflict of Interest: The 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.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 07 April 2021
doi: 10.3389/fcomp.2021.624594

Towards an Automatic Speech-Based Diagnostic Test for Alzheimer's Disease

Roozbeh Sadeghian[1]*, *J. David Schaffer*[2] *and Stephen A. Zahorian*[3]

[1]Department of Data Analytics, Harrisburg University, Harrisburg, PA, United States, [2]Institute for Justice and Well-Being, Binghamton University, Binghamton, NY, United States, [3]Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY, United States

Automatic Speech Recognition (ASR) is widely used in many applications and tools. Smartphones, video games, and cars are a few examples where people use ASR routinely and often daily. A less commonly used, but potentially very important arena for using ASR, is the health domain. For some people, the impact on life could be enormous. The goal of this work is to develop an easy-to-use, non-invasive, inexpensive speech-based diagnostic test for dementia that can easily be applied in a clinician's office or even at home. While considerable work has been published along these lines, increasing dramatically recently, it is primarily of theoretical value and not yet practical to apply. A large gap exists between current scientific understanding, and the creation of a diagnostic test for dementia. The aim of this paper is to bridge this gap between theory and practice by engineering a practical test. Experimental evidence suggests that strong discrimination between subjects with a diagnosis of probable Alzheimer's vs. matched normal controls can be achieved with a combination of acoustic features from speech, linguistic features extracted from a transcription of the speech, and results of a mini mental state exam. A fully automatic speech recognition system tuned for the speech-to-text aspect of this application, including automatic punctuation, is also described.

Keywords: speech processing, natural language processing, machine learning, alzheimer's disease, dementia

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Kevin El Haddad,
University of Mons, Belgium
Loredana Sundberg Cerrato,
Nuance Communications,
United States

Correspondence:
Roozbeh Sadeghian
rsadeghian@harrisburgu.edu

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science.

Received: *31 October 2020*
Accepted: *16 February 2021*
Published: *07 April 2021*

Citation:
Sadeghian R, Schaffer JD and Zahorian SA (2021) Towards an Automatic Speech-Based Diagnostic Test for Alzheimer's Disease. Front. Comput. Sci. 3:624594. doi: 10.3389/fcomp.2021.624594

INTRODUCTION

Dementia is broadly defined as deterioration in memory, thinking and behavior that decreases a person's ability to function independently in daily life (McKhann et al., 2011). The clinical diagnosis of dementia, particularly Alzheimer's disease (AD), is very challenging, especially in its early stages (Dubois et al., 2015). It is widely believed to be underdiagnosed, even in developed countries, and even more so in less developed countries. As people live longer, the prevalence of AD is huge and growing, with more than five million AD sufferers estimated in the US alone and an annual negative economic impact of over $200 billion (Association, 2019). New diagnostics are appearing, but they are often costly (e.g. involving brain imaging or novel lab tests), invasive (e.g. involving spinal taps, blood samples or the use of radioactive tracers), or both. A simple quick non-invasive test would be very desirable. In addition, recruitment for clinical trials of putative dementia therapies is hampered by lack of tests capable of yielding cohorts with a high likelihood of having the condition the therapy is designed to effect. An accurate diagnostic test would increase the feasibility and reliability of clinical outcome monitoring.

There are good indications that dementias can be characterized by several aphasias (defects in the production and use of speech) (Jacobs et al., 1995; Lowit et al., 2006; Cuetos et al., 2007). This seems plausible since speech production involves many brain regions, and thus a disease that effects particular regions seems likely to leave detectable "finger prints" in the speech of those with dementia. There are many relevant background scientific studies reported in the literature including those that attempt to establish specific voice-based features whose distributions are statistically different between those with dementia vs. normal controls (Bucks et al., 2000; Pakhomov et al., 2010; Meilán et al., 2014; König et al., 2015; López-de-Ipiña et al., 2015). Recently, a study by Eyigoz et al. (Manera et al., 2020) has provided additional evidence that the emergence of AD can be predicted using linguistic features.

Using speech as a neuropsychological assessment tool is now widely accepted. For example, the Boston Naming Test (BNT) (König et al., 2015) asks patients to see a picture and respond to questions within a short amount of time. Verbal fluency by describing a picture (Hernández-Domínguez et al., 2018) is another approach involved in diagnosing Alzheimer's. In most of these works, the features are manually extracted and their correlation to psychological benchmarks such as MCI (Mild Cognitive Impairment) or MMSE (Mini-Mental State Exam) are analyzed. MMSE, is a neuropsychological test (pencil and paper) which yields a score in the 0–30 range in about 10–15 min. Scores above 25 usually are assumed to indicate normal cognition. While not specifically designed for Alzheimer's diagnosis, it is often a first assessment applied by physicians, and can provide a useful "first cut" assessment. In the current work, we report on experimental results that combine MMSE test scores, basic demographic features (age, gender, race, and years of education) and a pool of features extracted from a voice sample. Using pattern discovery algorithms to identify minimal size feature sets, we provide evidence that combining selected speech features with the MMSE can yield an improved diagnostic test for detecting probable Alzheimer's disease. These results were obtained using features extracted automatically by algorithms applied to the speech signal (wave file) and either manually produced transcripts or fully automated transcripts produced by a custom designed ASR and punctuation system. The manually generated transcripts and automatically generated transcripts achieve approximately the same level of diagnostic precision, giving support to the hypothesis that current speech recognition technology is capable of supporting a fully automatic system.

The classical approaches to AD diagnosis (McKhann et al., 2011) rely on clinical criteria, often using neuropsychological tests, but require an autopsy for definitive diagnosis. Hence, recently, there has been much effort devoted to more reliable tests, seeking biomarkers in bodily fluids or imaging. Unfortunately, such tests are usually costly in time and money and bring their own risks (e.g. using radioactive tracers, or punctures). It is widely believed that AD is underdiagnosed, particularly in the undeveloped world, but also in the more developed nations. It is also believed that the disease pathology is at work years or decades before cognitive decline becomes apparent. This inability to accurately detect the disease early and accurately may have also contributed to the failures of clinical trials of putative AD agents. Hence, we believe there is a strong need for an accurate diagnostic test that is easy to execute, non-invasive and inexpensive. Here, we present results of efforts to produce such a test based on samples of human speech.

TABLE 1 | Demographic Summary of dataset.

Grp	*n*	Age (sd)	Years edu (sd)	MMSE (sd)
NL	46	71.43 (12.6)	13.28(2.4)	28.7 (1.5)
AD	26	78.48 (10.9)	13.81 (2.3)	20.92 (6.6)
Total	72	74.04 (12.4)	13.48 (2.4)	25.89 (5.6)

NL = Normal, AD = Alzheimer's disease; sd = Standard deviation, ed = education.

While there is yet no definitive evidence that such a test is possible, we subscribe to the intuition that speech, unique to our species, and requiring the coordinated activity of a number of brain regions, may have the characteristic that a lesion in one or more of these brain regions may well leave distinct finger prints in the speech. Furthermore, it is known that some speech-based tests have diagnostic utility (e.g. verbal fluency). Given recent advances in computational linguistics, this intuition seems to have a growing following based on the recent increase of research publications aimed in this direction.

We perceive two major challenges: one scientific, and one engineering. Can we provide convincing evidence that accurate diagnosis is possible with speech-based features, and can such a test be automated to the level that relatively untrained clinicians can use it? We believe our results provide encouragement that both challenges can be met.

This paper is organized as follows. In *Speech Sample Collection* we review the collected dataset and give some analysis of the dataset. In *Using ASR to Obtain the Transcripts*, the methodology of speech-to-text analysis is described. In *Classifier Design* the final machine learning model that we used is described. Conclusions are given in *Conclusions*.

SPEECH SAMPLE COLLECTION

A popular protocol for collecting speech samples for aphasia analysis work is to ask volunteers to describe what they see in a picture. They are able to view the picture while they speak. This protocol was used for all speech samples used in this work. There are some speech samples available on the web from the Dementia Bank audio database (Weiss and Bourgeois, 2012), but the audio quality is quite low. For our earlier work (Schaffer et al., 2005), we did exploit 140 of the Dementia Bank cases using manually prepared transcripts. This significantly increased our sample size. These samples were examined for use, but were generally of too low quality to be used for the experimental work reported in this study, especially the automatic speech recognition component. Since our long range goal was a fully automatic diagnostic tool, later work used our ASR system, which limited our samples to our own with high audio quality.

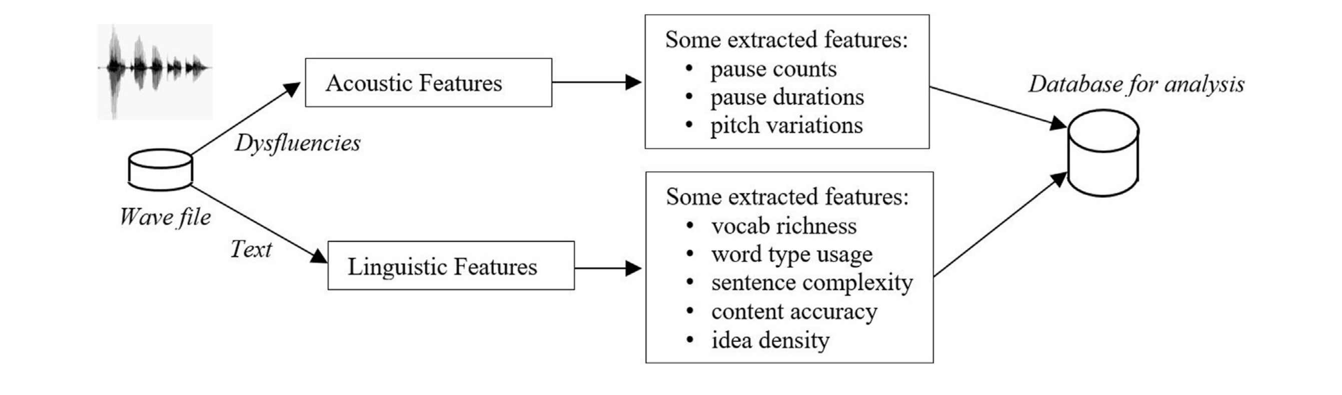


FIGURE 1 | Block diagram for creating database of features.

Since we elected not to use the Dementia bank database, 72 new samples (summarized in **Table 1**) were collected using modern digital recording equipment and a new picture (Sadeghian et al., 2015; Sadeghian et al., 2017). Twenty-six of these participants were AD (identified by measuring MMSE score, and verified by physician assessments) while 46 of them were normal. The average sample length was 75.1 s (sd 61.0 s) and the average age was 73.8 years (sd 12.1 yrs). Some modest preprocessing was performed on audio files, such as removing the beginning and ending pauses, click removal and signal strength normalization. These steps are straightforward to automate. The resulting acoustic speech files were processed directly for acoustic features such as pauses and pitch contours. A manual transcript was generated for each of the 72 samples. The manual transcripts were used to extract linguistic features (e.g. word counts, syntactic complexity, idea density). For comparison, we also created transcripts using ASR (Automatic Speech Recognition) and give diagnostic results based on the automatic methods.

Features, Subset Selection and Classification Approach

Each transcript was passed to the Charniak Parser (Charniak, 2000) trained with the Penn Treebank Switchboard corpus. The raw text of the transcript, and the part-of-speech (POS) tagged parser outputs were used to compute a number of linguistic metrics. These metrics include (but not limited to): average number of words per sentence, percentage of sentences that are classified as being "short," i.e. at most 5 words, length of the shortest sentence, the fraction of the words in the transcript that are auxiliary verbs or infinitives.

The syntactic complexity measures computed by Roark et al. (2011)) were computed, including a re-implementation of idea density (Snowdon et al., 1996). A number of metrics that capture various aspects of vocabulary richness were also computed as well as counts of words related to the picture content. The Linguistic Inquiry Word Count (LIWC2015 (Pennebaker et al., 2015)) features were also computed. These and all the other features, such as speech pauses and pitch features, were combined into a single feature vector for each subject. These 231 features from the speech samples were combined with four demographic features and the MMSE score to give 236 total potential features. This feature computational procedure is illustrated in **Figure 1**.

TABLE 2 | ASR Word Accuracy (%).

Model	Train = Test	Train ≠ Test	Train ≠ Test and VAD
Monophone	37.9	22.7	41.2
Triphone	85.2	27.6	48.2
DNN	89.2	42.7	65.7

USING ASR TO OBTAIN THE TRANSCRIPTS

In a fully automatic system, all the steps must be done automatically, including the crucial step of speech-to-text. There were about 72 min of data collected from participants. All the ASR work was done with Kaldi software (Povey et al., 2011). We made use of a combination of Hidden Markov Model (HMM) and Deep Neural Network (DNN) methods. In the beginning stages of this work we attempted to use a commercial off the shelf system, but did not find it suitable to be adapted for this application.

ASR Design

The first step for designing an ASR system is to prepare the dictionary (lexicon), which is a listing of all the words used in the language model, and the allowable pronunciations for each of these words. For the ASR acoustic models, we first created simple monophone models, then used those models to design triphone models, and finally implemented a DNN-based recognizer using the triphone models. All models were built using 39 Mel Frequency Cepstral Coefficient (MFCC) features, computed with 25 ms frames spaced apart by 10 ms. A Bigram language model was developed based on the manual transcriptions. For monophone models, 3-state HMMs with 64 mixtures were used whereas for triphone models, 500 tied states were modeled with 8,000 Gaussian mixtures.

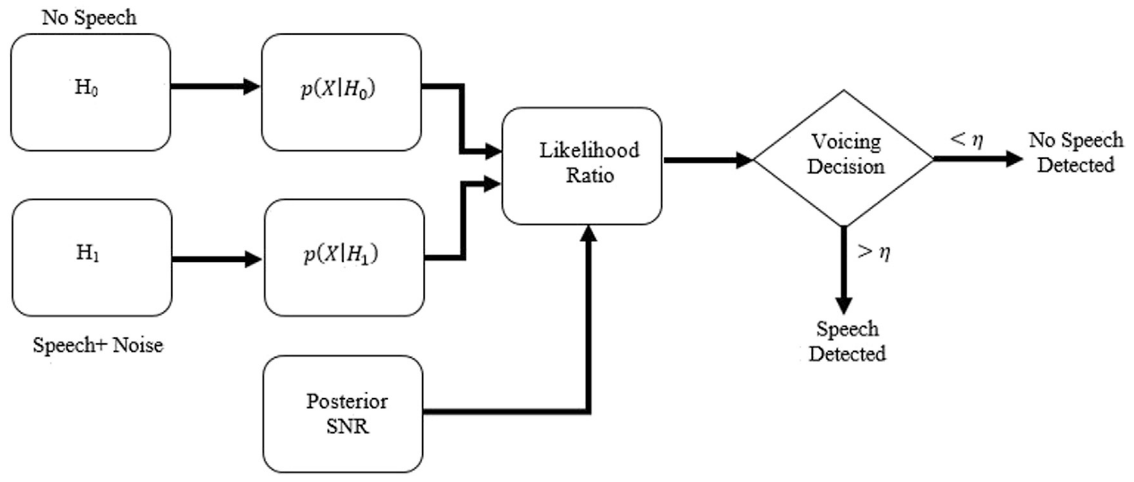


FIGURE 2 | Diagram of using VAD for speech Bishop, 2006.

For the DNN part of the recognizer, a network with two hidden layers was used in which each layer had 300 neurons. The initial learning rate was $\alpha = 0.015$ and it was decreased to $\alpha = 0.002$ in the final step. The activation function was hyperbolic tangent and the minibatch size was 128. To estimate the initial parameters of the model, we tested using the training data (ten different sets of test data were chosen with replacement from the same training cohort). Results are given in **Table 2**. We refer to cases where the train and test sets are the same as "cheating," and these cases are clearly not a true indication of performance on unseen data. Such cases, however, are useful to estimate an upper limit on accuracy possible with a given method.

When we used "honest" (completely separate training and test data) with 10-fold cross validation (Bishop, 2006), the word accuracy for test data was dramatically degraded to 47% from the best DNN case (89.2%) given in **Table 2**. This extremely poor generalization from training to test data led us to look for problems by carefully examining the speech files. By listening carefully to the files, we observed many silences (pauses) in the files that could be removed with an algorithm. We speculated that these silence intervals were severely degrading ASR accuracy. To address this issue we used a VAD (Voice Activity Detector) system.

A Voice Activity Detector (VAD) is a method for detecting the presence of speech in an audio signal. Several VAD algorithms are available (Savoji, 1989; Benyassine et al., 1997; Sohn et al., 1999). The method which we chose for this work was based on Sohn et al. (1999)). In this method, the unknown parameters are estimated using maximum likelihood (ML) and a likelihood Ratio Test (LRT). Further decision optimizations were performed using the decision directed (DD) method (Ephraim and Malah, 1984) and a hang-over scheme based on Hidden Markov Models (HMMs) for estimation of the unknown parameters. We describe and illustrate this method a little more in the next paragraph.

Consider a speech signal which is degraded by uncorrelated additive noise. In this case, for each frame we can define null and alternative hypotheses as (where S is signal and N represents noise):

$$\boldsymbol{H_0 : No\ Speech\ available : X = N}$$
$$\boldsymbol{H_1 : Speech\ available : X = N + S}$$

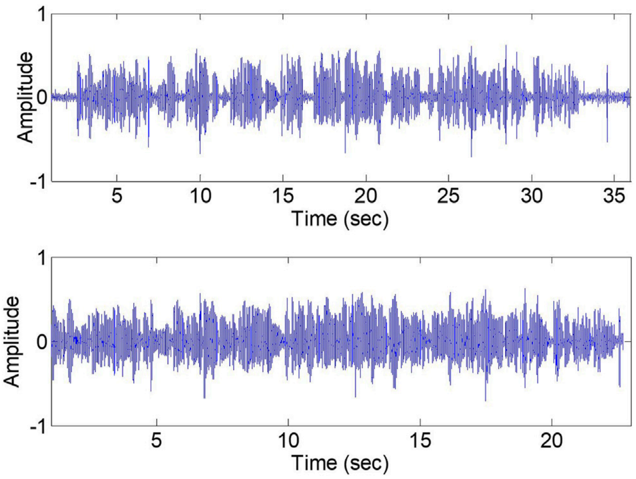


FIGURE 3 | Plot of sample speech segment before **(upper)** and after **(lower)** applying VAD.

An overview of the method is depicted in **Figure 2**. Applying VAD to the speech files removed an enormous amount of silence within the speech files. **Figure 3** shows the effect of this VAD on one of the speech samples from the database. As can be seen, most of the silence in the speech file is removed using the VAD algorithm. VAD helps to improve ASR accuracy. Although the average of the recognizer HMM + DNN "honest" accuracy is increased to around 65.7%, in comparison to state-of-the-art ASR methods, the accuracy still seems low.

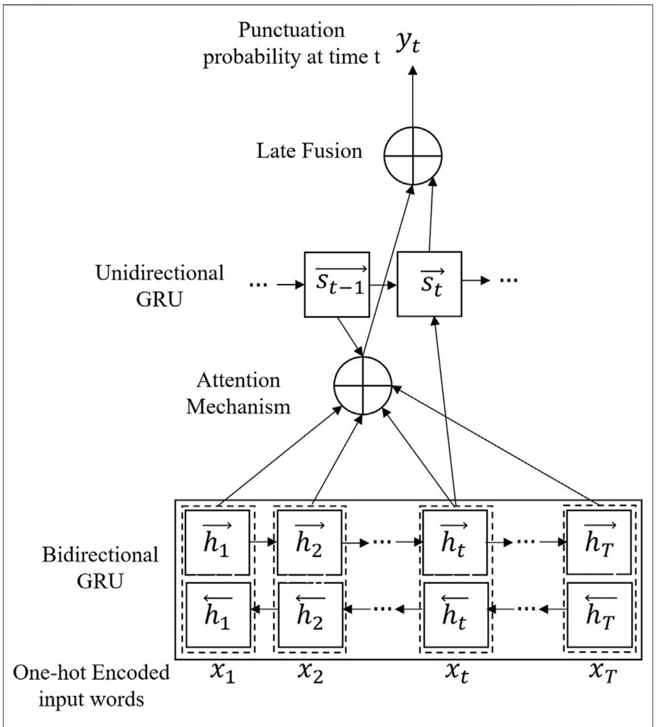

Punctuation probability at time t
y_t
Late Fusion
Unidirectional GRU
$\overrightarrow{s_{t-1}}$
$\overrightarrow{s_t}$
Attention Mechanism
Bidirectional GRU
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_t}$ $\overrightarrow{h_T}$
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_t}$ $\overleftarrow{h_T}$
One-hot Encoded input words
x_1 x_2 x_t x_T

FIGURE 4 | The structure of the punctuation detection model.

The accuracy highly depends on the number of speakers used for training. Increasing the size of the training database improves the acoustic model by using more samples which results in better Gaussian Mixture Model (GMM) and other parameter estimates. Additionally, the language model also highly depends on the number of training speakers. Since the number of speakers in the database was only 72, the best way to examine this effect was to use the Leave one out (LOO) method where we used 71 speakers for training and just one for testing, and then repeat for each speaker. The minimum accuracy among all the speakers was 22.4% and the maximum accuracy was 93.9%; the average accuracy was 68.7% with a standard deviation of 16. A closer examination of the worst case speaker revealed that the speaker still had a large number of pauses and OOV (Out of Vocabulary) words. Although the improvement in accuracy from 68% to 68.7% is very modest, at least it is in the right direction and also the LOO method allows us to look at the performance of each speaker individually. Based on the assumption that a word accuracy of 68.7% would be sufficient, and the fact that there were no clear-cut ways to improve this accuracy for this very difficult small database, we used the ASR method just described for the remainder of this work.

Automatic Punctuation

As mentioned earlier, we wanted to make use of two type of speech features, acoustic and linguistic features. For extracting linguistic features, the main punctuation needed is the sentence boundaries. The accuracy needed in the determination of sentence boundaries, the accuracy of determining "." vs. no punctuation, and the benefit of determining other punctuation is not clear. One method for automatic punctuation is to determine the sentence boundaries using a Support Vector Machine (SVM) and place the periods through a machine learning method (Beeferman et al., 1998). One other method which is popular is to use Conditional Random Fields (CRF) in the lexicon and, based on pause information, detect the sentence boundaries (Wei and Ng, 2010). Batista and Mamed (Batista and Mamede, 2011) used a combination of these two methods in Portuguese speech. The method that we used for this work was based on the method of Tilk (Tilk and Alum, 2016). In this method, a model based on a Recursive Neural Net (RNN) is developed which is trained using provided transcriptions. The structure of this model is depicted in **Figure 4**. The inputs of this model are one-hot encoded sequences of words in sentences where an end of sequence token is added to the list. The ultimate output of the network at time *t* is the prediction of the probability of punctuation y_t which is used between word $_{-1}$ and x_t. The Gated Recurrent Unit (GRU) approach was developed by Cho et al. (2014)) whereby each recurrent neuron captures the dependencies of different time scales adaptively. Using a GRU activation function with a shared embedding layer weight of W_e, the state h_t for the forward recurrent layer is defined by:

$$\overrightarrow{\boldsymbol{h}}_t = \mathbf{GRU}\left(\boldsymbol{x}_t \boldsymbol{W}_e, \overrightarrow{\boldsymbol{h}}_{t-1}\right).$$

Similarly, a reverse recurrent state can be defined with $\overleftarrow{h}_t$ whereby the words in the sentence sequence are processed in reverse. This type of configuration helps the model to identify if the sentence is in a declarative or question context. This means we assigned one layer of forward and one layer reverse recurrent state. Additionally, this allows the model to identify if a new sentence is started, considering the current word.

On top of this bidirectional state, there is a unidirectional GRU which keeps the track of the position at time t (based on the mechanism explained by Bahdanau et al. (2015)). There is a late attention that can consider both bidirectional and unidirectional outputs and creates an output to the late fusion step. The output of this model is the probability of using each punctuation at time t in the sequence of words. For our project, since only the boundary of sentences was important, we considered the "period" as the only punctuation that is required to be predicted. For this whole process of punctuation prediction, the effect of the acoustic part of speech is not considered. Tilk and Alum (2016) described another variation of this method in which another layer is added to the model that uses the effect of the duration of the pause in model design and it is considered part of the input training data. Although it may improve the results, this method was not used in this work, due to the added complexity.

For training the model, originally we used our own database but, because of the low number of sentences, the model was not accurate. Therefore, we used one of the available free databases, "Europal v7" (Kohen, 2005). In the English version of this corpus, 2,218,201 sentences from more than 800 speakers, containing more than 53 million words, were used. Around 90% of these data were used for training while 10% were used as a development (validation) set. There are two classes--no punctuation and

TABLE 3 | Confusion Matrix of Punctuation Detector.

Confusion Matrix		Actual values	
		Punctuation	No Punctuation
Predicted values	Punctuation	319	554
	No Punctuation	1048	8556

period. The training and testing data is chosen based on the sequence of the punctuations and their location in sentences.

The RNN model was trained using a learning rate of 0.02 while an L-2 norm of the gradient was kept below the threshold of 2 by renormalization whenever it exceeded the threshold. The stopping criteria for training was whenever the perplexity of the development set became worse for the first time during the iterations. In the first step, the weights were initialized using the normalization technique with zero bias. All the hidden layers contained 256 neurons. For training the model the Theano package (Bastien et al., 2012) with a GPU (Graphical Processing Unit) was used. The sequence of input words was chunked into 200 word long slices where each slice starts with the first word of the relevant sentence and if the slice ends with an unfinished sentence, the sentence is copied to the beginning of the next slice. Clearly, the output sequence is one element shorter since no punctuation is placed before the first word. Because of the huge amount of training data, the slices were grouped to mini-batches of 128 slices and were shuffled before each epoch. The output vocabulary can predict any punctuation such as comma, period, question mark and no punctuation. However, for this project, we only predicted the period. The error rate of punctuation prediction in this case is 15.3%. This error rate is computed by comparing the predicted punctuation and the actual one from manual punctuation. The f1-score (test of accuracy) was also obtained. This value is computed from the combination of precision (correctly positive predictive values out of all the predicted positive values) and recall (correctly positive predictive values out of all the actual positive values). The f1-score is below what we expected but this is mostly due to many OOVs in transcription that the DNN is not capable of punctuating accurately. The confusion matrix, cumulative over all 72 subjects, is given in **Table 3**.

CLASSIFIER DESIGN

The end goal of this work, from a technical perspective, is a two-way classifier to determine AD vs. NL (Normal) from a slate of features selected from the very large group described above. This problem is very challenging due to the very large number of candidate features (236), and the small database (72 speakers). We hypothesized that only 5–10 of these 236 features would be needed and useful for the final decision making. The challenge was to "discover" these "good" features using a small database, and in a manner that these features and classifier would perform well for data other than those used in this study. The two-way classifier model and the feature subset selector are depicted in **Figure 5**. All subjects were divided 90/10 into training and validation sets and full 10-fold cross validation was performed. For thoroughness, three approaches were used to investigate this step—GA-SVM, Random Forest, and Neural Network.

In our first approach, based on the diagram shown in **Figure 5**, a genetic algorithm (GA) was used as the feature subset selector while a Support Vector Machine (SVM) was the classifier which was trained using the features selected by the GA. This GA-SVM approach has been successfully applied to a number of bioinformatics pattern discovery tasks (Schaffer et al., 2005). This approach may generate and test more than a million subsets before it halts. An array of top candidates usually yields several alternative feature-set classifiers with differing performance. Summarizing, the genetic algorithm strives to locate feature sets of high accuracy and minimum size.

We used a 10-fold cross-validation approach. Each fold used 90% of subjects for feature subset identification and model training. The remaining subjects were tested only once. The different folds often found different feature sets to be best, but there were many commonalities. In the end, we identified 12 different feature sets, all comprising combinations of only 10 features. The five most important features are listed in **Table 4**. Each of these feature-sets (called classifiers) was then trained on each fold's training data and tested on the test cases only once. Experiments showed that 12 different classifiers typically made errors on different subjects, so their classification predictions could profitably be combined with an ensemble method. A Generalized Regression Neural Network (GRNN) (Specht, 1991) oracle is a maximum-likelihood, minimum variance unbiased estimator that has been shown to give very robust classification performance (Masters et al., 1998). Theoretically, it is the best one can do with a fixed data set. Since each of the 12 classifiers made different errors, they were combined using the GRNN oracle ensemble method yielding a single diagnosis predictor.

Since we had so few cases, we did not try to locate new features subsets that might have benefited from the automatic transcripts, but simply applied the same classifiers found on the large dataset, but tuned them separately for the manual and automatic transcripts in the same 10-fold fashion. The MMSE alone, the oracle using manual transcripts, and the oracle using automatic transcripts all made eight errors (8/72 = 11.1%), with seven subjects being misclassified on all. They made an error on one unique subject each. These seven common erroneously classified subjects were also errors of the oracle trained on the large dataset. From these results, we draw confidence that the fully automatic diagnostic test is likely to have the same success[1].

Random Forest

Random forest (RF) is a machine learning technique in which a decision tree is developed using the training data. RF was introduced by Ho (1995). Generally, decision trees or recursive partitioning models are a decision tool based on tree-like graphs

[1] The interested reader may find many more details on our analyses of these data in Walker and Schaffer (2019)

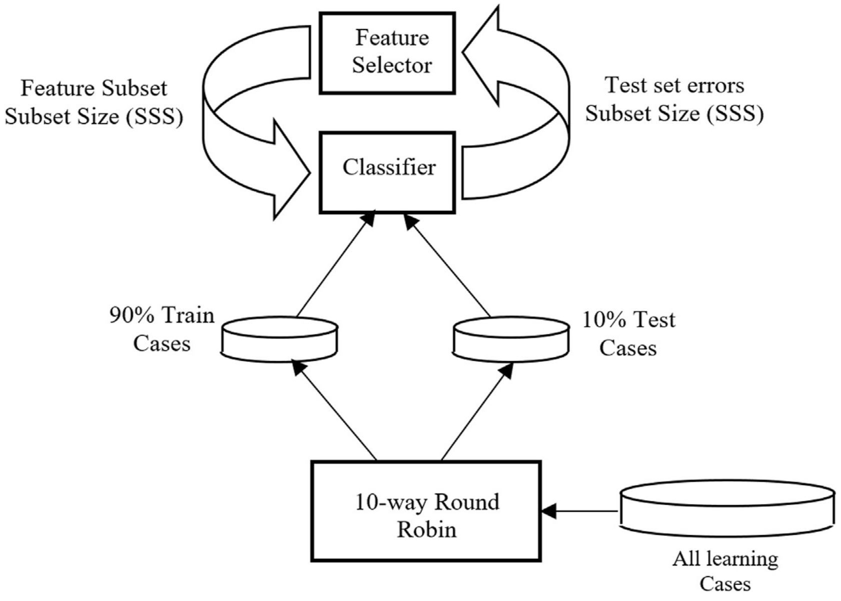


FIGURE 5 | Overview of classifier to determine AD/NL decision.

TABLE 4 | Selected Features By Classifier.

Feature subset (classifier)												Feature long name
1	2	3	4	5	6	7	8	9	10	11	12	
x	X	x	x	x	X	x	x	x	x	x	x	Mini-mental state exam score
	X											Fraction of the total utterance length that is speech (i.e. not pauses) (VAD based)
x	X	x	x	x	X	x	x	x	x	x		Fraction of utterance in pauses < 0.5 s (energy based)
x								x				Words > 6 letters Pennebaker et al. (2015)
x	X			x	x		x					Adjectives Pennebaker et al. (2015)
x	X				x	x	x	x	x	x	x	Male reference words Pennebaker et al. (2015)
	X						x			x		Special email words (e.g. BTW, LOL, emogies) and convenience words (e.g. ha, hm, huh, kinda, ya, yah, yup)
						x						Content density, the ratio of open-class words to closed-class words Roark et al. (2011)
	X											Readability score that estimates the United States. grade level necessary to understand a text
	X		x				x		x			Average syllables per word

and their possible consequences. It creates a flow chart which contains nodes (leaves) and decisions extracted from each node (branches). These leaves and branches form a tree-shaped graph which is referred to as decision tree. This model can yield high accuracy, robust performance and ease of use. This method is, however, highly sensitive to data. Hence, resampling is used to mitigate this issue. Every node in a decision tree, represents a decision (target) based on a single feature and a threshold which splits the dataset into two so that similar response values are collected in the same set. On this fully automated system, 90% of the data (66 subjects) were used for training while the remaining speakers were used just one time for testing. A 12-fold cross-validation was used with the training data which means that 90% of the data were used for training and the rest for a validation set. The training data were processed by the RF model to determine the best combination of the features. The total number of features was experimentally found to be 25, with10 trees in the forest. The function of the quality of split is called "Gini." Gini impurity is the factor showing how often a randomly selected label is incorrectly assigned based on the actual target distribution. Mathematically, it is the summation of the multiplication of the probability of the properly chosen label ($p_{i)}$) times the probability of the incorrectly chosen label ($1-p_i$) for all labels $i \in \{1, \ldots, K\}$. In equation form, the Gini impurity $I_G(p)$ is defined as:

$$\boldsymbol{I_G}(\boldsymbol{p}) = \sum_{\boldsymbol{i}=1}^{\boldsymbol{K}} \boldsymbol{p_i}(1 - \boldsymbol{p_i}).$$

The function is minimized when all the classes in the node lead to a similar target. Nodes were expanded until all leaves are pure or until all leaves contain fewer than two samples. Each node was split until its impurity was higher than a threshold of 1e-7; otherwise it was considered as a leaf. Due to the randomness of the process, the experiments were repeated 100 times and the most repeated features were considered as the desired ones. After finding the best combination of 25 features, these features were

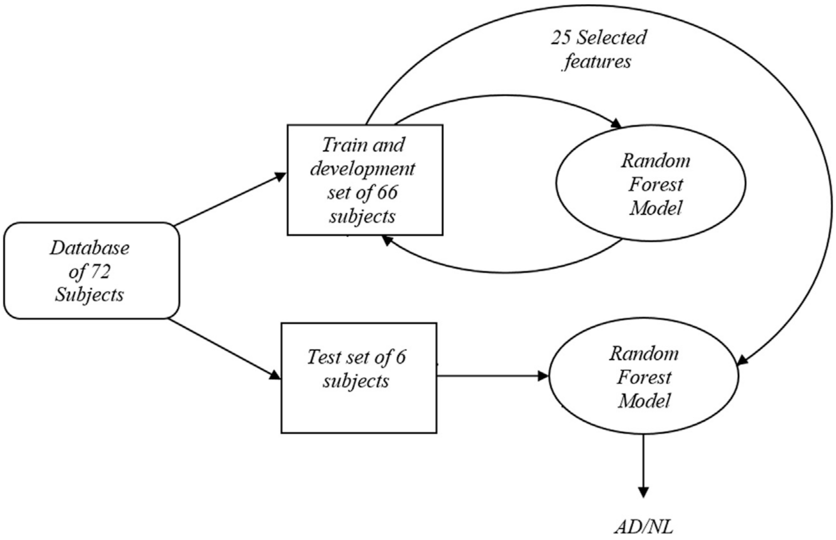


FIGURE 6 | Block diagram of random forest approach.

tested only one time on the test set. A diagram of this method is depicted in **Figure 6**.

A 12-fold cross validation test was considered whereby each speaker was used only one time as the test speaker. The overall accuracy using this methodology was 84.00%. Using LOO (Leave One Out) cross validation, improved the overall accuracy to 87.5% which shows again how having more data can improve the overall accuracy for this technique. In this technique, all speakers except one were used for feature selection and, after training the model using the most frequently used features, the model is evaluated on only one speaker. This procedure is repeated for all the subjects individually.

Multi-Layer Perceptron

As yet another method, the feature selection was repeated using a using a NN (Neural Network) - Multi-Layer Perceptron. For this model, a NN with one hidden layer (containing 25 nodes) was used as a two-way classifier. The activation functions were sigmoid. The inputs were features to be evaluated (from training data) and the outputs were assigned labels for each subject. A greedy approach was used whereby initially each of the 236 potential features was evaluated individually and the best performing feature was found. Best performance was determined by highest accuracy on a group of test speakers. After the best 1-feature classifier was found, the best 2-feature classifier was found by testing all 2-feature options, given that that one of these 2 features was the best feature for the 1-feature classifier. This process was repeated until some termination point (explained below) was reached.

The initial experiments "over fit" the training set due to minimizing the expected loss instead of empirical loss defined on the training set. To resolve this issue, a weight decay (Krogh and Hertz, 1991) term was added to the loss function, i. e.

$$\frac{\partial^2 \boldsymbol{l}}{\partial \boldsymbol{W}^2}(\boldsymbol{W}, \boldsymbol{b}) = \boldsymbol{l}(\boldsymbol{W}, \boldsymbol{b}) + \boldsymbol{\lambda R}(\boldsymbol{W}),$$

where *l(W,b)* is the original loss function, λ is the weight decay parameter and *R(W)* is defined by:

$$\mathbf{R}(\mathbf{W}) = \|\mathbf{vec}(\mathbf{W})\|.$$

The decay parameter (L1-regularization) for this experiments was set to be 0.1 experimentally. Another popular approach for preventing overfitting, which was also used in our work, was dropout (Srivastava et al., 2014). The idea of dropout is that α percent of neurons are omitted from hidden layers during the training phase. This adds some random noise to the network through some hidden layers whereas even with similar inputs, there is no guarantee that higher layers will receive similar inputs. This is achieved by forcing the activation nodes to zero while in the test phase the average of the neurons are used. The rate of dropout for this work was 0.02, again experimentally determined.

For inputting data to the model, the Stochastic Gradient Descent (SGD) (Bottou, 1998) technique was used. This method updates the parameters of the NN model from only a single training sample. One main advantage of SGD is that, despite batch learning, due to its noisy gradient estimation it can easily jump out of the local minima in estimation iterations.

For the first part of the experiments, 72 subjects (fully automated system) were used where the data were partitioned as explained previously. Ten-fold cross validation was used to find the best combination of the features through the greedy approach described above. The best feature sets which were revealed by validation data were later used on a test set to determine the accuracy of the model. Using these features, the average accuracy of the testing set was 94.44%. As a comparison, the same model was created using

TABLE 5 | Final Accuracy, Sensitivity, and Specificity (%) of Different ML Models.

Algorithm	Accuracy	Sensitivity	Specificity
Random Forest- KFold	84.0	77.8	91.1
Random Forest- LOO	87.5	74.1	95.6
ANN- Fully automated	94.4	92.3	95.7
ANN- Manual transcript	95.8	96.2	95.7
ANN- MMSE feature only	70.8	88.5	60.9

manual transcripts, which increased the accuracy slightly to 95.83%. The top three features using this strategy were MMSE, fraction of pauses greater than 1 s and fraction of total speech recording that was silence.

As we mentioned earlier, the MMSE score is the most important feature (best single feature) for an AD/NL classifier. To see its strength, we tested this feature as the only feature for the classifier. An accuracy of 70.83% was obtained. To see how well the MMSE feature could be compensated for by other features, we removed the MMSE feature and used the NN feature selector/classifier as described above. This gave an accuracy of 91.67% for manual transcripts and 93.05% for the automatic one. The top three features using this strategy were speech rate, idea density and fraction of pauses greater than 1 s.

A summary of the key results for the random forest and neural network two way classifiers (AD/NL) are given in **Table 5**, in terms of accuracy, sensitivity, and specificity.

CONCLUSIONS

There do appear to be strong patterns among the speech features that are able to discriminate the subjects with probable Alzheimer's disease from the normal controls. The GA-based feature subset selection approach provides a powerful way to locate multiple classifiers that contain many common features combined with some less common ones, lending themselves to being combined with ensemble methods (Masters et al., 1998). We have shown this elsewhere (Land and Schaffer, 2015) along with a method for enabling the classifier to know when it should not be trusted. However, these results are likely to be sensitive to small samples, suggesting larger samples should be used for future research in this domain.

The greedy algorithm combined with the neural network two-way classifier was very promising for both feature selection and final recognizer. For feature selection, this approach was at least two orders of magnitude faster than the GA method. The limitation of the NN method is that the search of the feature space is not nearly as exhaustive as for the GA method. In future work, the NN method could be improved in terms of more thorough searching by saving the top N (where N is some small number such as 5–10) choices at the end of each iteration, at the expense of some slowdown in speed. The NN classifier, using common features, was also as good as the SVM used as the final classifier with the GA.

We believe this study provides encouragement to seek speech patterns that could be diagnostic for dementia. The weaknesses of this study, aside from the obviously very small sample size, include the cross-sectional design that strives for a single pattern that works over the whole variety of subjects in each class. A longitudinal study would permit each subject to serve as his own control, helping to mitigate the large within-group variance in speaking patterns, as well as introduce the possibility for predicting dementia that is currently not manifest. The features used are by no means all the speech features that have been associated with dementia. The computational linguistics domain contains several additional interesting speech features that, with some effort, could be included in our basket of candidate features.

The best accuracy of ~ 96% achieved in this study for diagnosing Alzheimer seems promising considering the small number of samples used. Additionally, the results of manually and automatically transcribed systems are similar, which shows that the ASR system worked in an acceptable range and the punctuator system was likely accurate enough. Summarizing across all the Alzheimer's experiments, we conclude the following with respect to features (from possibilities including MMSE score, demographics, and acoustic speech features, linguistic speech features) and approximate detection accuracy:

1) The most informative single parameter is the MMSE alone, which results in a detection accuracy of about 71%.
2) If all possible features, including MMSE scores, are considered, a detection accuracy of approximately 94% is possible, using fully automatic methods. Based on the features listed in **Table 4**, MMSE is always chosen as one of the key features. Three linguistic and one acoustic features are selected, which are fraction of pauses more than 5 s in duration, speech and the LIWC quantitative feature.
3) If all possible features, except MMSE scores, are considered, a detection accuracy of approximately 92% is possible, based on the features listed in **Table 4**.
4) If only demographic and acoustic features (the "easy" ones) are considered, a detection accuracy of approximately 83.33% is possible. However, for this case, there was low sensitivity. That is, there was a high error rate for AD subjects (often diagnosed as NL). The most important features for this case are speech rate (using energy and VAD), fraction of speech length to the length of whole audio and fraction of pause length to the whole audio file.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because we will consider sharing the data with others with the consent of the University's Ethics Committee, as it may always be possible to identify research subjects from their voices. Requests to access the datasets should be directed to David Schaffer, dschaffe@binghamton.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the State University of New York at Binghamton University and United Health Services. All subjects provided written informed consent, and in the cases of subjects with AD, a healthcare proxy also provided informed consent. The patients/participants provided their written informed consent to participate in this study.

REFERENCES

Association (2019). 2015 Alzheimer's disease facts and figures. *Alzheimers Dement* 11, 332–384. doi:10.1016/j.jalz.2015.02.003

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv*.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., et al. (2012). Theano: new features and speed improvements. *arXiv*.

Batista, F., and Mamede, N. (2011). Recovering capitalization and punctuation marks on speech transcriptions. PhD dissertation. Lisboa, Portugal: Instituto Superior Técnico.

Beeferman, D., Berger, A., and Lafferty, J. (1998). "Cyberpunc: a lightweight punctuation annotation system for speech," in Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98, Seattle, WA, May 15, 1998 (IEEE). doi:10.1109/ICASSP.1998.675358

Benyassine, A., Shlomot, E., Su, H. Y., Massaloux, D., Lamblin, D., and Petit, J. P. (1997). ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Commun. Mag.* 35, 64–73. doi:10.1109/35.620527

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin, Germany: Springer-Verlag.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line Learn. Neural networks* 17 (9), 142. doi:10.1017/CBO9780511569920.003

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology* 14 (1), 71–91. doi:10.1080/026870300401603

Cho, K., Merrienboer, B. V., Gulcehre, C., Bahdanau, D., Bougares, F., Schewenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv*. doi:10.3115/v1/D14-1179

Cuetos, F., Arango-Lasprilla, J. C., Uribe, C., Valencia, C. F., and Lopera, F. (2007). Linguistic changes in verbal expression: a preclinical marker of alzheimer's disease. *J. Int. Neuropsy. Soc.* 13, 433–439. doi:10.1017/S1355617707070609

Dubois, B., Padovani, A., Scheltens, P., Rossi, A., and Dell'Agnello, G. (2015). Timely diagnosis for alzheimer's disease: a literature review on benefits and challenges. *J. Alzheimers Dis.* 49 (3), 617–631. doi:10.3233/jad-150692

Charniak, E. (2000). "A maximum-entropy-inspired parser," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, April, 2000, Stroudsburg, PA, United States, pp. 132–139.

Ephraim, Y., and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* 32 (6), 1109–1121. doi:10.1109/tassp.1984.1164453

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., and Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's Demen. Assess. Dis. Monit.* 10, 260–268. doi:10.1016/j.dadm.2018.02.004

Ho, T. K. (1995). "Random decision forests," in Proceedings of the 3rd international conference on document analysis and recognition, Montreal, QC, August 14–16, 1995 (IEEE), 278–282. doi:10.1109/ICDAR.1995.598994

Jacobs, D. M., Sano, M., Dooneief, G., Marder, K., Bell, K. L., and Stern, Y. (1995). Neuropsychological detection and characterization of preclinical alzheimer's disease. *Neurology* 45, 957–962. doi:10.1212/wnl.45.5.957

AUTHOR CONTRIBUTIONS

RS worked on implementing the speech recognition and machine learning tests and providing the first draft of paper. JS, collected the samples and created the manual transcripts and worked on initial GA-SVM model. He also worked on preparing the manuscript. SZ, was the lead of team and advisor of the group. He worked on preparing the final manuscript.

Kohen, P. (2005). A parallel corpus for statistical machine translation. *MT summit* 5, 79–86.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst)* 1, 112–124. doi:10.1016/j.dadm.2014.11.012

Krogh, A., and Hertz, J. A. (1991). "A simple weight decay can improve generalization," in Proceedings of the 4th International Conference on Neural Information Processing Systems, San Francisco, CA, December 1991 (NIPS), 950–957.

Land, W. H., and Schaffer, J. D. (2015). Predicting with confidence: extensions to the GRNN oracle enabling quantification of confidence in predictions. *Proced. Comp. Sci.* 61, 381–387. doi:10.1016/j.procs.2015.09.164

López-de-Ipiña, K., Solé-Casals, J., Eguiraun, H., Alonso, J. B., Travieso, C. M., Ezeiza, A., et al. (2015). Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: a fractal dimension approach. *Comp. Speech Lang.* 30 (1), 43–60. doi:10.1016/j.csl.2014.08.002

Lowit, A., Dobinson, B. C., and Howell, P. (2006). An investigation into the influences of age, pathology and cognition on speech production. *J. Med. Speech Lang. Pathol.* 14, 253–262.

Manera, E., Mathur, S., and Santamaria, M. (2020). Guillermo cecchi, and melissa naylor, "linguistic markers predict onset of alzheimer's disease. *Eclinical Med.* 27, 100583. doi:10.1016/j.eclinm.2020.100583

Masters, T., Land, W. H., and Maniccam, S. (1998). "An oracle based on the general regression neural network," in IEEE International Conference on Systems, Man, and Cybernetics - SMC, San Francisco, CA, October 14, 1998 (IEEE), 1615–1618.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., et al. (2011). The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimers Dement* 7, 263–269. doi:10.1016/j.jalz.2011.03.005

Meilán, J. J., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia?. *Dement Geriatr. Cogn. Disord.* 37 (5–6), 327–334. doi:10.1159/000356726

Pakhomov, S. V., Smith, G. E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., et al. (2010). Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cogn. Behav. Neurol.* 23 (3), 165–177. doi:10.1097/WNN.0b013e3181c5dde3

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: The University of Texas at Austin.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). "The Kaldi speech recognition toolkit," in IEEE 2011 workshop on automatic speech recognition and understanding, Hawaii, United States, December 2011 (IEEE).

Roark, B., Mitchell, M., Hosom, J. P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans. Audio Speech Lang. Process.* 19 (7), 2081. doi:10.1109/TASL.2011.2112351

Sadeghian, R., Schaffer, D. J., and Zahorian, S. A. (2015). Using automatic speech recognition to identify dementia in early stages. *The J. Acoust. Soc. America* 138 (3), 1782. doi:10.1121/1.4933648

Sadeghian, R., Schaffer, J. D., and Zahorian, S. A. (2017). "Speech processing approach for diagnosing dementia in an early stage", in InterspeechStockholm, August 20–24 2017, Sweden. doi:10.21437/interspeech.2017-1712

Savoji, M. H. (1989). Robust algorithm for accurate end pointing of speech. Amesterdam, Netherlands: Speech communication.

Schaffer, J. D., Janevski, A., and Simpson, M. (2005). "Genetic algorithm approach for discovering diagnostic patterns in molecular measurement data," in IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, CA, November 15, 2005 (IEEE), 7803–9387. doi:10.1109/CIBCB.2005.1594945

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. *JAMA J. Amer. Medi. Ass.* 275 (7), 528–532. doi:10.1001/jama.275.7.528

Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal. Process. Lett.* 6 (1), 1–3. doi:10.1109/97.736233

Specht, D. F. (1991). A general regression neural network. *IEEE Trans. Neural Netw.* 2 (6), 568–576. doi:10.1109/72.97934

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Machine Learn. Res.* 15 (1), 1929–1958.

Tilk, O., and Alum, T. (2016). "Bidirectional recurrent neural network with attention mechanism for punctuation restoration", in Interspeech, September 8–12 2016, San Fransisco. doi:10.21437/interspeech.2016-1517

Walker, H. J., and Schaffer, J. D. (2019). *The art and science of machine intelligence: with an innovative application for Alzheimer's detection from speech*. Switzerland: Springer, International Publishing AG.

Wei, L., and Ng, H. T. (2010). "Better punctuation prediction with dynamic conditional random fields", in Proceedings of the 2010 conference on empirical methods in natural language processing, Massachusetts, October 9–10, 2010.

Weiss, J., and Bourgeois, M. (2012). Development of DementiaBank: language usage in dementia. *Presented at American speech-language-hearing association convention*. Atlanta, GA.

Conflict of Interest: The 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.

ORIGINAL RESEARCH
published: 08 April 2021
doi: 10.3389/fcomp.2021.624694

Longitudinal Speech Biomarkers for Automated Alzheimer's Detection

Jordi Laguarta[1] *and Brian Subirana*[1,2]*

[1] *MIT AutoID Laboratory, Cambridge, MA, United States,* [2] *Faculty of Arts and Sciences, Harvard University, Cambridge, MA, United States*

We introduce a novel audio processing architecture, the Open Voice Brain Model (OVBM), improving detection accuracy for Alzheimer's (AD) longitudinal discrimination from spontaneous speech. We also outline the OVBM design methodology leading us to such architecture, which in general can incorporate multimodal biomarkers and target simultaneously several diseases and other AI tasks. Key in our methodology is the use of multiple biomarkers complementing each other, and when two of them uniquely identify different subjects in a target disease we say they are orthogonal. We illustrate the OBVM design methodology by introducing sixteen biomarkers, three of which are orthogonal, demonstrating simultaneous above state-of-the-art discrimination for two apparently unrelated diseases such as AD and COVID-19. Depending on the context, throughout the paper we use OVBM indistinctly to refer to the specific architecture or to the broader design methodology. Inspired by research conducted at the MIT Center for Brain Minds and Machines (CBMM), OVBM combines biomarker implementations of the four modules of intelligence: The brain OS chunks and overlaps audio samples and aggregates biomarker features from the sensory stream and cognitive core creating a multi-modal graph neural network of symbolic compositional models for the target task. In this paper we apply the OVBM design methodology to the automated diagnostic of Alzheimer's Dementia (AD) patients, achieving above state-of-the-art accuracy of 93.8% using only raw audio, while extracting a personalized subject saliency map designed to longitudinally track relative disease progression using multiple biomarkers, 16 in the reported AD task. The ultimate aim is to help medical practice by detecting onset and treatment impact so that intervention options can be longitudinally tested. Using the OBVM design methodology, we introduce a novel lung and respiratory tract biomarker created using 200,000+ cough samples to pre-train a model discriminating cough cultural origin. Transfer Learning is subsequently used to incorporate features from this model into various other biomarker-based OVBM architectures. This biomarker yields consistent improvements in AD detection in all the starting OBVM biomarker architecture combinations we tried. This cough dataset sets a new benchmark as the largest audio health dataset with 30,000+ subjects participating in April 2020, demonstrating for the first time cough cultural bias.

Keywords: multimodal deep learning, transfer learning, explainable speech recognition, brain model, graph neural-networks, AI diagnostics

OPEN ACCESS

Edited by:
Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:
Mary Jean Amon,
University of Central Florida,
United States
Andrea Seveso,
University of Milano-Bicocca, Italy

****Correspondence:***
Brian Subirana
subirana@mit.edu

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *01 November 2020*
Accepted: *25 February 2021*
Published: *08 April 2021*

Citation:
Laguarta J and Subirana B (2021)
Longitudinal Speech Biomarkers for
Automated Alzheimer's Detection.
Front. Comput. Sci. 3:624694.
doi: 10.3389/fcomp.2021.624694

1. INTRODUCTION

Since 2001, the overall mortality for Alzheimer's Dementia (AD) has been increasing year-on-year. Between 2000 and 2020 deaths resulting from stroke, HIV and heart disease decreased while reported deaths from AD increased by about 150% (Alzheimer's Association, 2020). Currently no treatments are available to cure AD, however, if detected early on, treatments may greatly slow and eventually possibly even halt further deterioration (Briggs et al., 2016).

Currently, methods for diagnosing AD often include neuroimaging such as MRI (Fuller et al., 2019), PET scans of the brain (Ding et al., 2019), or invasive lumbar puncture to test cerebrospinal fluid (Shaw et al., 2009). These diagnostics are far too expensive for large-scale testing and are usually used once family members or personal care detect late-stage symptoms, when the disease is too advanced for onset treatment. On top of the throughput limitations, recent studies on the success of the most widely used form of diagnostic, PET amyloid brain scans, have shown expert doctors in AD currently misdiagnose patients in about 83% of cases and change their management and treatment of patients nearly 70% of the time (James et al., 2020). This is mainly caused by the lack of longitudinal explainability of these scans. As a result it is hard to track effectiveness of treatments and even more to evaluate personalized treatments tailored to specific on-set populations of AD (Maclin et al., 2019). AI in general suffers from similar issues and operates a bit as a black-box, and does not offer explainable results linked to specific causes of each individual subject (Holzinger et al., 2019).

Based on the above findings, our research aims to find AD diagnostic methods achieving the following four warrants:

1. **Onset Detection:** detection needs to occur as soon as the first signs emerge, or sooner even if only probabilistic metrics can be provided. Preclinical AD diagnosis and subsequent treatment may offer the best chances at delaying the effects of dementia (Briggs et al., 2016). Therapeutic significance may require establishing subclassifications within AD (Briggs et al., 2016). Evidence that there are early signs of AD onset in the human body come in the form of recent research on blood plasma phosphorylated-tau isoforms diagnostic biomarkers demonstrating chemical traces of dementia, and of AD in particular, decades in advance of clinical diagnosis (Barthélemy et al., 2020; Palmqvist et al., 2020). These are encouraging findings, and hopefully there are also early onset signs in free-speech audio signals. In fact, preclinical AD is often linked to mood changes and in cognitively normal adults onset AD includes depression (Babulal et al., 2016), while apathy and anxiety have been linked to some cognitive decline (Bidzan and Bidzan, 2014). Both of these may be detectable in preclinical AD using existing sentiment analysis techniques (Zunic et al., 2020).
2. **Minimal Cost:** we need a method that has very little side effects, so that a person can perform the test periodically, and at very low variable costs to allow broad pre-screening possibilities. Our suggestion is to develop methods that can run on smart speakers and mobile phones (Subirana et al., 2017b) at essentially no cost while respecting user privacy (Subirana et al., 2020a). There is no medically approved system allowing preclinical AD diagnosis at scale. There are different approaches to measure AD disease onset and progression but all rely on expensive human assessments and/or medical procedures. We demonstrate our approach using only free speech but the approach can also include multi-modal data if available including MRI images (Altinkaya et al., 2020) and EEG recordings (Cassani et al., 2018).
3. **Longitudinal tracking:** the method should include some form of AD degree metric, especially to evaluate improvements resulting from medical interventions. The finer disease progression increments can be measured, the more useful they'll be. Ideally, adaptive clinical trials would be supported (Coffey and Kairalla, 2008).
4. **Explainability:** the results need to have some form of explainability, if possible including the ability to diagnose other types of dementia and health conditions. Most importantly, the approach needs to be approved for broad use by the medical community.

Our approach is enabled by and improves upon advances in deep learning on acoustic signals to detect discriminating features between AD and non-AD subjects—it aims to address the warrants above, including explainability which has been challenging for previous approaches. While research in AD detection from speech has been ongoing for several years most approaches did not surpass the 90% detection mark as shown in **Table 1**. These approaches use black-box deep learning algorithms providing little to no explainability as to what led the model's decision, making it hard for clinicians to use and hence slowing adoption by the healthcare system. In Petti et al. (2020), review of the literature on AD speech detection, about two thirds of the papers reviewed use Neural Nets or Support Vector Machines, while the rest focus on Decision Trees and Naïve Bayes. Neural Nets seem to achieve the highest detection accuracy on average. Previous work, instead, has very little inspiration on the different stages of human intelligence and at most focuses solely on modeling a small part of the brain as shown in Nassif et al. (2019), de la Fuente Garcia et al. (2020), and Petti et al. (2020).

TABLE 1 | A review of other AD diagnostic algorithms on the same dataset from Lyu (2018).

References	Date	Accuracy(%)
Syed et al. (2020)	2020	85.4
Haulcy and Glass (2021)	2021	85.4
Orimaye et al. (2014)	2016	87.5
Yuan et al. (2020)	2020	89.6
Karlekar et al. (2018)	2018	91.1
Laguarta and Subirana	2021	93.8

Our top performing model only uses audios while Orimaye et al. only used 35 patients hence risking high variance. Karlekar et al. only used transcripts. The rest used the transcripts from the ADreSS challenge Luz et al. (2020).

Combining independent biomarkers with recent advances in our understanding of the four modules of the human brain as researched at MIT's Center for Brain Minds and Machines (CBMM) (CBM, 2020), we introduce a novel multi-modal processing framework, the MIT CBMM Open Voice Brain Model (OVBM). The approach described in this paper aims to overcome limitations of previous approaches, firstly by training the model on large speech datasets and using transfer learning so that the accurate learned features improve AD detection accuracy even if the sample of AD patients is not large. Secondly, by providing an explainable output in the form of a saliency chart that may be used to track the evolution of AD biomarkers.

The use of independent biomarkers in the CBMM Open Voice Brain Model enables researching what is the value of each of them, simply by contrasting results with and without one of the biomarkers—we illustrate this point with a biomarker focused on cough discrimination (Subirana et al., 2020b) and one focused on wake words (Subirana, 2020). We feel this is an original contribution of our work grounded on the connection between respiratory conditions and Alzheimer's.

Furthermore, we also show that our framework lets apply the same biomarker models for audio detection of multiple diseases, and explore whether there may be common biomarkers between AD and other diseases. To that end, the OVBM framework we introduce may be extended to various other tasks such as speech segmentation and transcription. It has already proven to detect COVID-19 from a forced-cough recording with high sensitivity including 100% asymptomatic detection (Laguarta et al., 2020). Here we demonstrate it in the individualized and explainable diagnostic of Alzheimer's Dementia (AD) patients, where, as shown in **Table 1** we achieve above state-of-the-art accuracy of 93.8% (Pulido et al., 2020), and using only raw audio as input, while extracting for each subject a saliency map with the relative disease progression of 16 biomarkers. Even with expensive CT scans, to date experts can not create consistent biomarkers as described in James et al. (2020), Henriksen et al. (2014), and Morsy and Trippier (2019) even when including emotional biomarkers, unlike our approach which automatically develops them from free speech. Experts point at this lack of biomarkers as the reason why no new drug has been introduced in the last 16 years despite AD (Zetterberg, 2019) being the sixth leading cause of death in the United States (Alzheimer's Association, 2019), and one of the leading unavoidable causes for loss of healthy life.

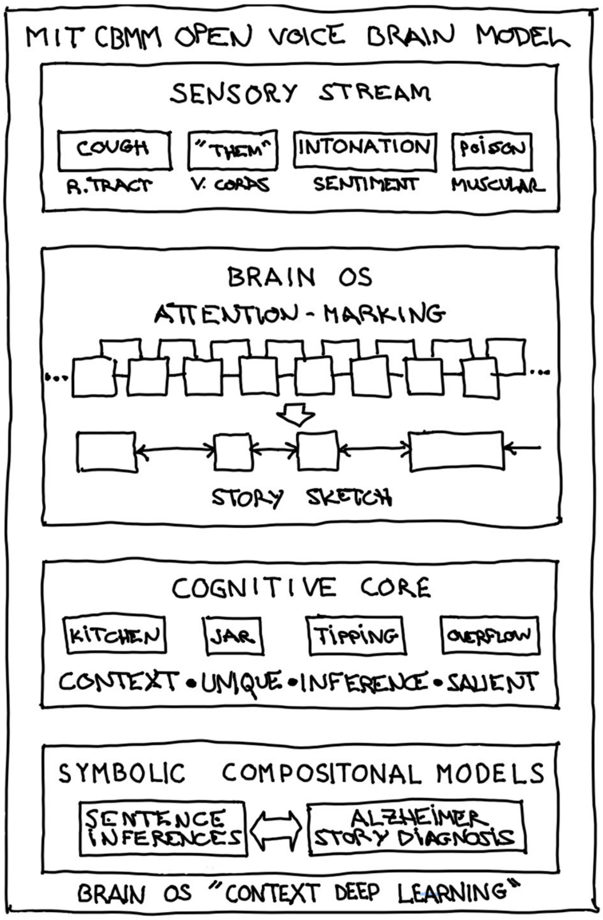


FIGURE 1 | Diagram of MIT CBMM open voice 4 module brain model with the selected AD Biomarkers.

We found that cough features, in particular, are very useful biomarker enablers as shown in several experiments reported in this paper and that the same biomarkers could be used for COVID-19 and AD detection. Our emphasis on detecting relevant biomarkers corresponding to the different stages of disease onset, led us to build ten sub-models using four datasets. To do so, over 200,000 cough samples were crowd sourced to pre-train a model discriminating English from Catalan coughs, and then transfer learning was leveraged to exploit resulting features by integrating it into an OVBM brain model, showing improvements in AD detection, no matter what transfer learning strategy was used. This COVID-19 cough dataset we created approved by MIT's IRB 2004000133 sets a new benchmark as the largest audio health dataset, with over 30,000 subjects participating in less than four weeks in April 2020.

In the next section we present a literature review with evidence in favor of our choice of four biomarkers. In section 3, we present the different components of the Open Voice Brain Model AD detector, from sections 4 to 7 we introduce the 16 biomarkers with results and a novel personalized AD biomarker comparative saliency map. We conclude in section 8 with a brief summary and implications for future research.

2. LITERATURE REVIEW SUPPORTING OUR CHOICE OF FOUR SENSORY STREAM AUDIO BIOMARKERS: COUGH, WAKE WORD, SENTIMENT, AND MEMORY

Informed by a review of the literature, our choice of biomarkers is consistent with the vast literature resulting from AD research as we discuss next.

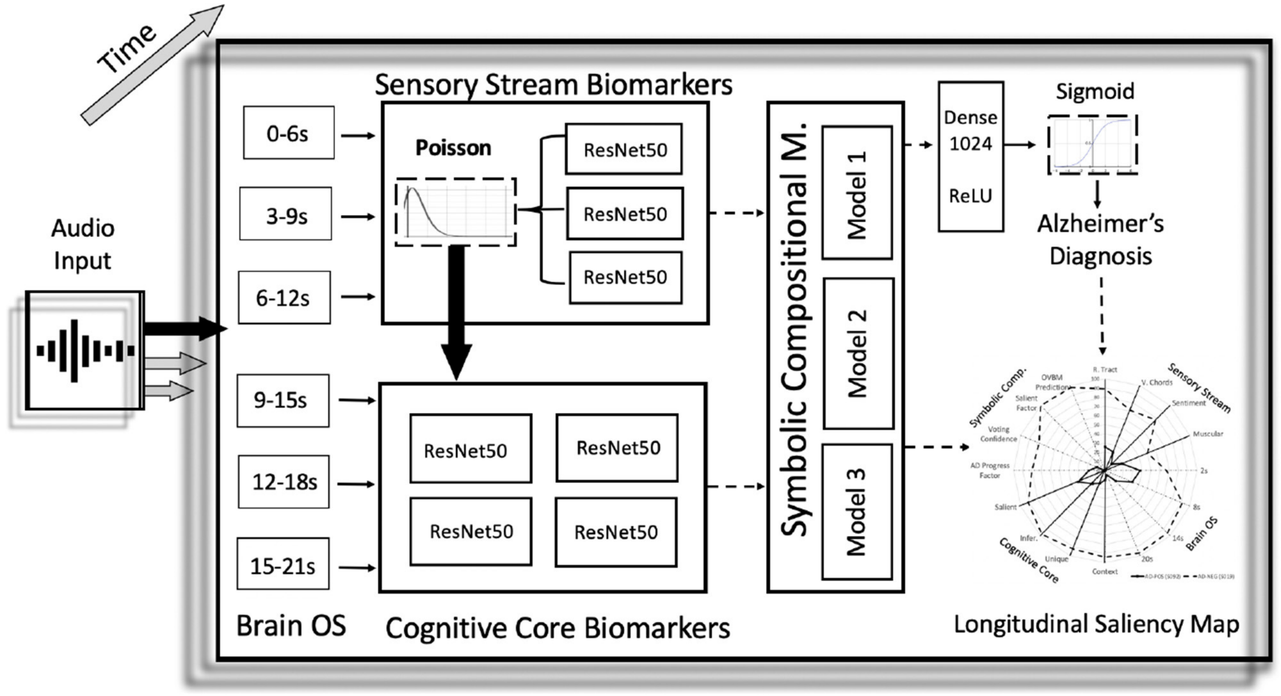


FIGURE 2 | OVBM GNN architecture at a given Brain OS time.

2.1. Mood Biomarkers

Preclinical AD is often linked to mood changes. In cognitively normal adults it include depression (Babulal et al., 2016), while apathy and anxiety have been linked to some cognitive decline (Bidzan and Bidzan, 2014). Sentiment biomarker. Clinical evidence supports the importance of sentiments in AD early-diagnosis (Costa et al., 2017; Galvin et al., 2020), and different clinical settings emphasize different sentiments, such as doubt, or frustration (Baldwin and Farias, 2009).

2.2. Memory Biomarkers

One of the main early-stage AD biomarkers is memory loss (Chertkow and Bub, 1990), which occurs both at a conceptual level as well as at a muscular level (Wirths and Bayer, 2008) and is different from memory forgetting in healthy humans (Cano-Cordoba et al., 2017; Subirana et al., 2017a). A prominent symptom of early stage AD is malfunctioning of different parts of memory depending on the particular patient (Small et al., 2000), possibly affecting one or more of its subcomponents including primary or working memory, remote memory, and semantic memory. The underlying causes of these memory symptoms may be linked to neuropathological changes, such as tangles and plaques, initially affecting selected areas of the brain like the hippocampi or the temporal and frontal lobes, and gradually expanding beyond these (Morris and Kopelman, 1986). Memory biomarker.

2.3. Respiratory Tract Biomarkers Cough and Wake Word

The human cough is already used to diagnose several diseases using audio recognition (Abeyratne et al., 2013; Pramono et al., 2016) as it provides information corresponding to biomarkers in the lungs and respiratory tract (Bennett et al., 2010). People with chronic lung disorders are more than twice as likely to have AD (Dodd, 2015), therefore we hypothesize features extracted from a cough classifier could be valuable for AD diagnosis.

There is an extensive cough-based diagnosis research of respiratory diseases but to our knowledge, no one had applied it to discriminate other, apparently unrelated, diseases like Alzheimer's. Our findings are consistent with the notion that AD patients cough differently and that cough-based features can help AD diagnosis; they are also consistent with the notion that cough features may help detect the onset of the disease. The lack of longitudinal datasets prevents us from exploring this point but do allow us to demonstrate the diagnostic power of cough-based features, to the point where without these features we would not have surpassed state-of-the-art performance.

The respiratory tract is often involved in the fatal outcome of AD. We introduce two biomarkers focused on the respiratory tract that may help discriminate between early and late stage AD. We have not found research indicating how early changes in the tract may be detected but given it's importance in the disease outcome it may be early on. This could also explain the success of many speech-based AD discrimination approaches—some of which have been applied to early stages of FTD. Significant research in AD such as Heckman et al. (2017)

TABLE 2 | Impact of Poisson mask on AD performance.

Model	W/o Poisson(%)	With Poisson(%)
Baseline	65.6	68.8
Cough	75.0	75.0
Intonation	68.8	75.0
Wake-Word "Them"	75.0	78.1
Multi-Modal	90.6	93.8
Avg improvement(%)		3.1

Baseline is a ResNet50 trained on the AD task without transfer learning.

TABLE 3 | To illustrate the complementary nature of the biomarkers we show the unique AD patients detected by each individual biomarker model with only the final classification layer fine-tuned on the target disease, Alzheimer's and COVID-19 in this case.

Biomarker	Model Name	Alzheimer's(%)	COVID-19(%)
Respiratory tract	Cough	9	23
Sentiment	Intonation	19	8
Vocal cords	WW "THEM"	16	19
R. Tract and sentiment	Cough and Tone.	0	0
R. Tract and vocal cords	Cough and WW	6	1
Sentiment and vocal cords	Tone. and WW	3	0
In all 3		41	34
In neither of the 3		6	15

Each transfer model detects unique patients reinforcing orthogonality of the biomarkers and hence the potential of combining new ones. Note how exactly the same biomarker models can detect Alzheimer's and COVID-19 subjects, showing the transferable nature for different diseases and how they behave "orthogonally" in both cases.

has proven that the disease impacts motor neurons. In other diseases, like Parkinson's, where motor neurons are affected, vocal cords have proven to be one of the first muscles affected (Holmes et al., 2000).

Dementia in general has been linked to increased deaths from pneumonia (Wise, 2016; Manabe et al., 2019) and COVID-19 (Azarpazhooh et al., 2020; Hariyanto et al., 2020) possibly linked to specific gens (Kuo et al., 2020). COVID-19 deaths are more likely with Alzheimer's than with Parkinson's disease (Yu et al., 2020). This different respiratory response depending on the type of dementia suggests that related audio features, such as coughs, may be useful not only to discriminate dementia subjects from others but also to discriminate specific types of dementia.

We contend there is correlation, instead of causality, between our two respiratory track biomarkers and Alzheimer's but further elucidation to this extent is necessary as there is in many other areas with AD and more broadly in science in general (Pearl and Mackenzie, 2018). Some causality link may exist due to the simultaneous role of substance P in Alzheimer's (Severini et al., 2016) and in cough (Sekizawa et al., 1996). The existence of spontaneous cough *per se* may not be enough to predict onset risk but in combination with other health parameters may contribute to an accurate risk predictor (Song et al., 2011). Our biomarker suggestion is based on "forced coughs" which, to our knowledge, has not been studied in connection with Alzheimer's. We feel it may be an early indication of future respiratory tract conditions that will show in the form of spontaneous coughs. In patients with late-onset Alzheimer's Disease (LOAD) a unique delayed cough response has been reported in COVID-19 infected subjects (Isaia et al., 2020; Guinjoan, 2021). Dysphagia and aspiration pneumonia continue to be the two most serious conditions in late stage AD with the latter being the most common cause of death of AD patients (Kalia, 2003), suggesting substance P induced early signs in the respiratory tract may already be present in forced coughs, perhaps even unavoidably.

What seems unquestionable is the connection between speech and orofacial apraxia and Alzheimer's, and it has been suggested that it, alone, can be a good metric for longitudinal assessment (Cera et al., 2013). Various forms of apraxia have been linked to AD progression in different parts of the brain (Giannakopoulos et al., 1998). Nevertheless, given the difficulty in estimating speech and orofacial apraxia these figures are not part of common Clinical Dementia Rating scales (Folstein et al., 1975; Hughes et al., 1982; Clark and Ewbank, 1996; Lambon Ralph et al., 2003). However, all these studies reveal difficulties in an objective, accurate, and personalized scale that can track each patient independently from the others (Olde Rikkert et al., 2011). The lack of metrics also spans other related indicators such as quality of life estimations (Bowling et al., 2015). There are no reliable biomarkers for other neurogenerative disorders either (Johnen and Bertoux, 2019).

Recent research has demonstrated that apraxia screening can also predict dementia disease progression (Pawlowski et al., 2019), especially as a way to predict AD in early stage FTD subjects, a population that we are particularly interested in targeting with our biomarkers. For the Behavioral Variant of Fronto Temporal Dementia (bvFTD), in patients under 65 the second most common cognitive disorder caused by neurodegeneration, little tonal modulation and buccofacial apraxia, are targeted by our biomarkers and are established diagnostic domains (Johnen and Bertoux, 2019). We hope that our research can help establish reliable biomarkers for disease progression that can also distinguish at onset between the different possible diagnostics. The exact connection between buccofacial apraxia and dementia has not been as well-documented as that of other forms of apraxia. Recent results show that there buccofacial apraxia may be present in up to fifty percent of dementia patients with no association to oropharyngeal dysphagia (Michel et al., 2020). Oropharyngeal dysphagia, on the other hand, has been linked to dementia, in some studies in over fifty percent of the cases, appearing, in particular, in late stages of FTD and in early stages of AD (Alagiakrishnan et al., 2013).

According to the NIH's National Institute of Neurological Disorders and Stroke information page on apraxia[1], the most common form of apraxia is orofacial apraxia which causes the inability to carry out facial movements on request such as coughing. Cough reflex sensitivity and urge-to-cough deterioration has been shown to help distinguish AD from

[1] https://www.ninds.nih.gov/disorders/all-disorders/apraxia-information-page.

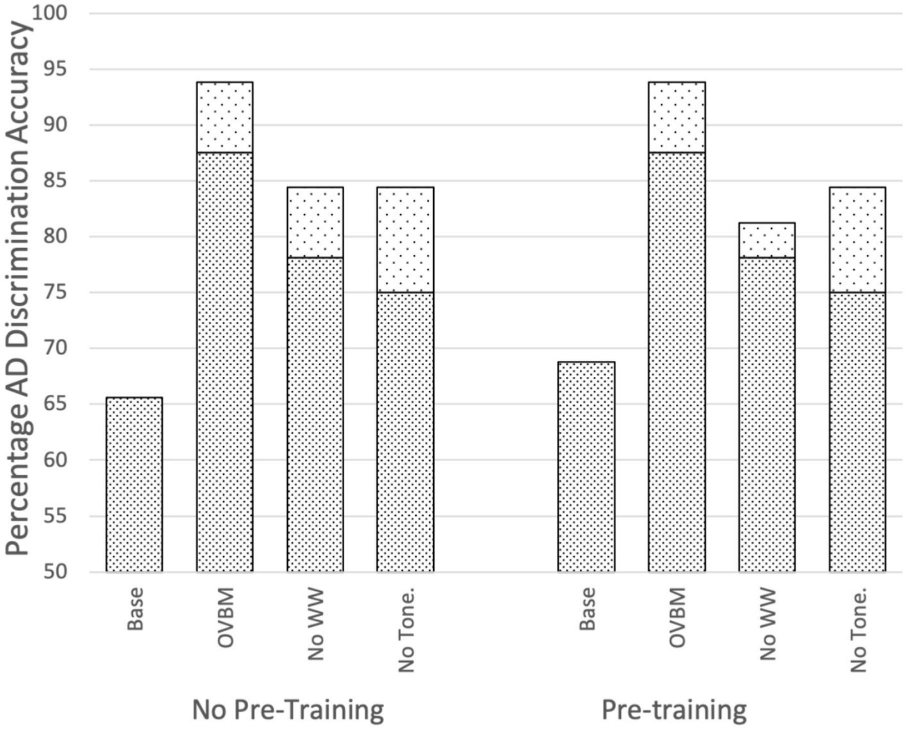


FIGURE 3 | Impact of sensory stream biomarkers on OVBM performance by removing transferred knowledge one at a time. Top dotted sections of bars indicate there is always performance gain from the cough biomarker. Baselines are the OVBM trained on AD without any transfer learning. In the other bars, a biomarker is removed and replaced with an AD pre-trained ResNet50, hence removing the transferred knowledge but conserving computational power, showing complementarities since all are needed for maximum results.

dementia with Lewy Bodies and control groups (Ebihara et al., 2020). The impairment of cough in the elderly is linked to dementia (Won et al., 2018).

3. OVERVIEW OF THE MIT OPEN VOICE BRAIN MODEL (OVBM) FRAMEWORK

The OVBM architecture shown in **Figure 1** frames a four-unit system to test biomarker combinations and provides the basis for an explainable diagnostic framework for a target task such as AD discrimination. The Sensory Stream is responsible for pre-training models on large speech datasets to extract features of individual physical biomarkers. The Brain OS splits audio into overlapping chunks and leverages transfer learning strategies to fine-tune the biomarker models to the smaller target dataset. For longitudinal diagnosis, it includes a round-robin five stage graph neural network that marks salient events in continuous speech. The Cognitive Core incorporates medical knowledge specific to the target task to train cognitive biomarker feature extractors. The Symbolic Compositional Models unit combines fine-tuned biomarker models into a graph neural network. Its predictions on individual audio chunks are fed into an aggregating engine subsequently reaching a final diagnostic plus a patient saliency map. To enable doctors to gain insight into the specific condition of a given patient, one of the novelties of our approach is that the outputs at each unique module are extracted to create a visualization in the form of a health diagnostic saliency map showing the impact of the selected biomarkers. This saliency map may be used to longitudinally track and visualize disease progression.

3.1. OVBM Applied to AD Detection

Next, we review each of the four OVBM modules in the context of AD, introducing 16 biomarkers and gradually explaining the partial GNN architecture shown in **Figure 2**. To be able to compare models, our baselines and 8 of the biomarkers are based on the ResNet50 CNN due to its state-of-the-art performance on medical speech recognition tasks (Ghoniem, 2019). All audio samples are processed with the MFCC package published by Lyons et al. (2020), and padded accordingly. We operate on Mel Frequency Cepstral Coefficients (MFCC), instead of spectrograms (Lee et al., 2009), because of its resemblance to how the human cochlea captures sound (Krijnders and t Holt, 2017). All audio data uses the same MFCC parameters (Window Length: 20 ms, Window Step: 10 ms, Cepstrum Dimension: 200, Number of Filters: 200, FFT Size: 2,048, Sample rate: 16,000). All datasets follow a 70/30 train-test split and models are trained with an Adam optimizer (Kingma and Ba, 2014).

The dataset from DementiaBank, ADrESS (Luz et al., 2020), is used for training the OVBM framework and fine-tuning all biomarker models on AD detection. This dataset is the

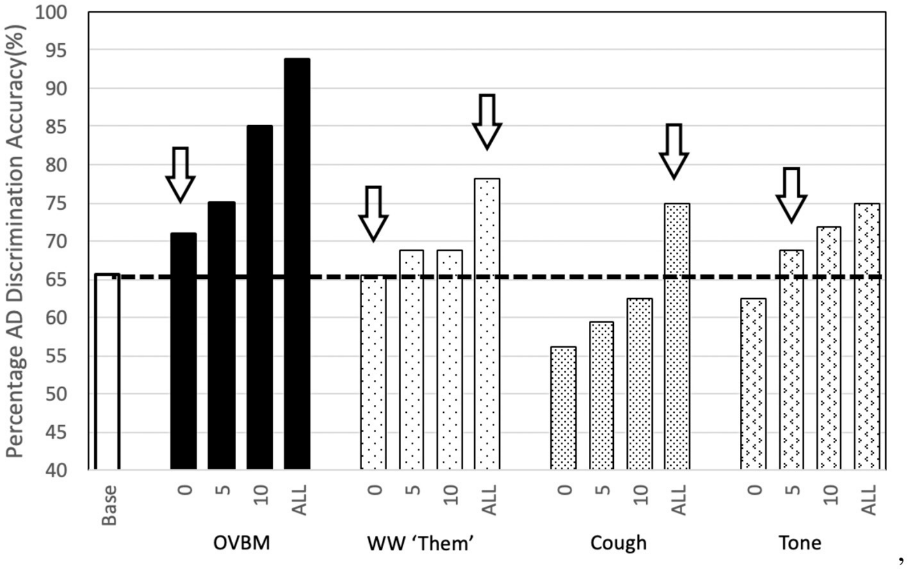


FIGURE 4 | Sensory Stream Saliency Bar Chart: To illustrate the potential of our approach we show the strength of the simplest transfer models we tried. The numbers 0-5-10-ALL on the x-axis labels refer to the number of convolution layers trained after transfer learning in addition to the final dense layer. We find the most surprising, perhaps, is that the simple wakeword model to find the word "Them" is as powerful as the baseline. If we let the model fine-tune the last few (0-5-10) layers then it goes well beyond it. Our novel Cough database, inspired in the effect of AD in the respiratory tract also shows surprising results, even without any adaptation at all. If we let fine-tuning of the whole model, it's validation accuracy improves ≈10% points with respect to the baseline. Baseline is the same OVBM architecture trained on AD without any transfer learning of features.

largest publicly available, consisting of subject recordings in full enhanced audio and short normalized sub-chunks, along with the recording transcriptions from 78 AD and 78 non-AD patients. The patient age and gender distribution is balanced and equal for AD and non-AD patients. For the approach of this study focusing purely on audio processing we only use the full enhanced audio and patient metadata, excluding transcripts from any processing. It is worth noting this given the poor audio quality of some of the recordings.

4. OVBM AD SENSORY STREAM BIOMARKERS

We have selected four biomarkers inspired by previous medical community choices (Chertkow and Bub, 1990; Wirths and Bayer, 2008; Dodd, 2015; Heckman et al., 2017; Galvin et al., 2020), as reviewed next.

4.1. Biomarker 1 (Muscular Degradation)

We follow memory decay models from Subirana et al. (2017a) and Cano-Cordoba et al. (2017) to capture this muscular metric by degrading input signals for all train and test sets with the **Poisson** mask shown in Equation (1), a commonly occurring distribution in nature (Reed and Hughes, 2002). We use as a Possion function a mask with input MFCC image = I_x, output mask = $M(I_X)$, $\lambda = 1$, and k = each value in I_x:

$$M(I_x) = Pr(\lambda)I_x \tag{1}$$

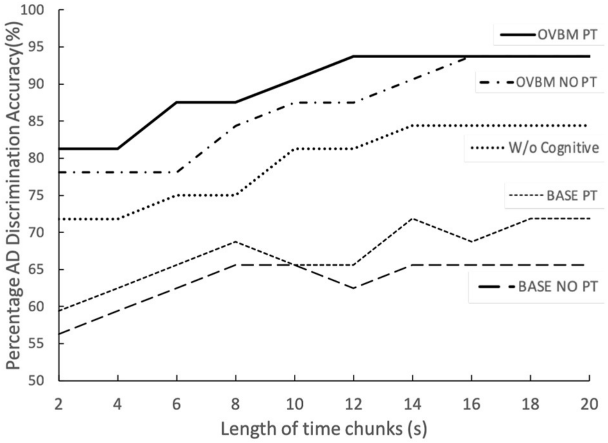


FIGURE 5 | The two top lines illustrate the full OVBM performance, with its biomarker feature models, as a function of chunk size. PT refers to individually fine-tuning each biomarker model for AD before re-training the whole OVBM. The middle line shows the OVBM without the cognitive core, illustrating how it boosts performance by about 10% across the board. Baseline PT is the OVBM architecture with each ResNet50 inside individually trained on AD before retraining them together in the OVBM architecture.

$$Pr(X = k) = \frac{\lambda^k e^{-k}}{k!}$$

As shown in **Table 2**, this Poisson biomarker brings a unique improvement to each model except for Cough, consistent with both inherently capturing similar features containing muscular degradation.

4.2. Biomarker 2 (Vocal Cords)

We have developed a vocal cord biomarker to incorporate in OBVM architectures. We trained a Wake Word (WW) model to learn vocal cord features on LibriSpeech—an audiobook dataset with ≈1,000 h of speech from Panayotov et al. (2015) by creating a balanced sample set of 2 s audio chunks, half containing the word "Them" and half without. A ResNet50 (He et al., 2016) is trained for binary classification of **"Them"** on 3 s audio chunks(lr:0.001, val_acc: 89%).

Illustrated in **Table 3** and **Figure 4**, this vocal cords model proves its contribution of unique features, which without fine-tuning to the AD task performs as well as the baseline ResNet50 fully trained on AD, and significantly beats it when fully fine-tuned.

4.3. Biomarker 3 (Sentiment)

We train a Sentiment Speech classifier model to learn **intonation** features on RAVDESS—an emotional speech dataset by Livingstone and Russo (2018) of actors speaking in eight different emotional states. A ResNet50 (He et al., 2016) is trained on categorical classification of eight corresponding intonations such as calm, happy, or disgust(lr: 0.0001, val_acc on 8 classes: 71%).

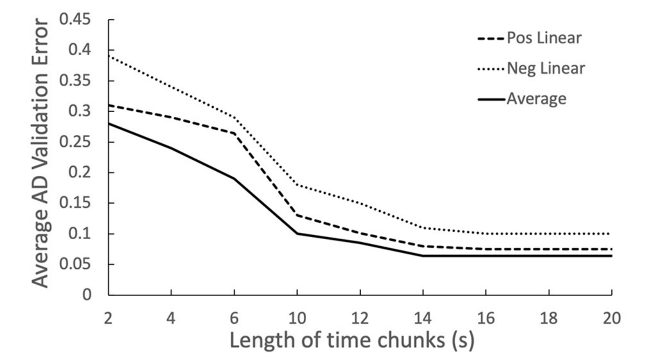


FIGURE 6 | Relation between chunk size and AD discrimination error, showing increased importance of the latter chunks.

As illustrated by **Table 3** and **Figure 4**, this biomarker captures unique features for AD detection, and when only fine-tuning its final five layers outperforms a fully trained ResNet50 on AD detection.

4.4. Biomarker 4 (Lungs and Respiratory Tract)

We use the **cough** dataset collected through MIT Open Voice for COVID-19 detection (Subirana et al., 2020b), strip all but the spoken language of the person coughing (English, Catalan), and split audios into 6 s chunks. A ResNet50 (He et al., 2016) is trained on binary classification (Input: MFCC 6s Audio Chunks (1 cough)—Output: English/Catalan, lr: 0.0001, val_acc: 86%).

Figure 4 and **Table 3**, justify the features extracted by this cough model as valuable for the task of AD detection by capturing a unique set of samples and improving performance. Further, **Figure 3** validates its impact on various OBVM architectures, including the top performing multi-modal model, justifying the relevance of this novel biomarker.

5. OVBM BRAIN OS BIOMARKERS

The Brain OS is responsible for capturing learned features from the individual biomarker models in the Sensory Stream and Cognitive Core, and for integrating them into an OVBM architecture, with the aim of training the ensemble for a target task, in this case AD detection.

To make the most out of the short patient recordings, we split each patient recording into overlapping audio chunks (0–4, 2–6, 4–8 s). Once the best pre-trained biomarker models in

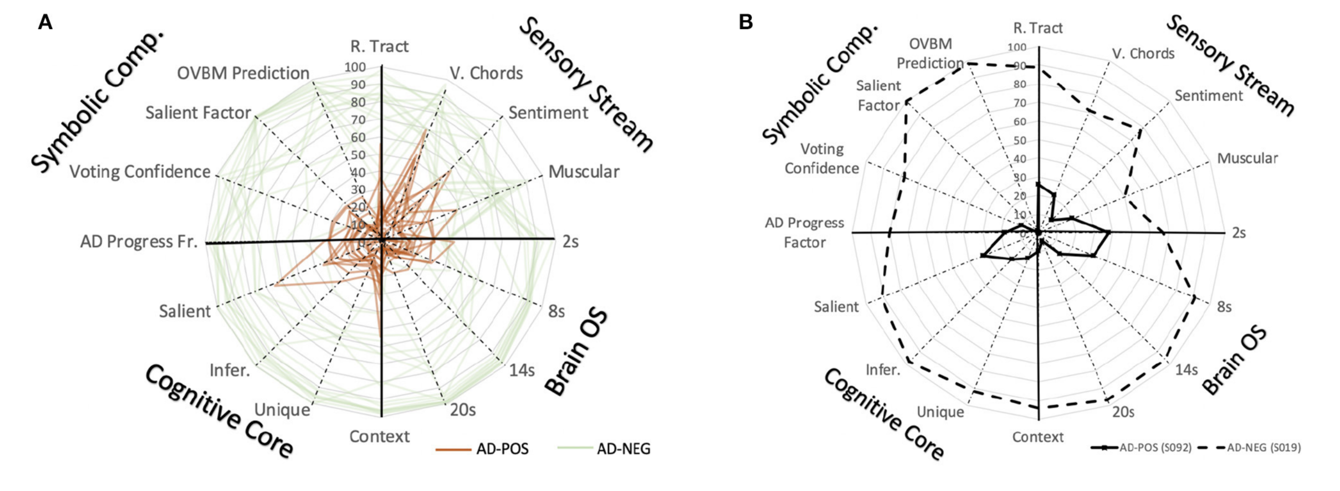


FIGURE 7 | (A) Saliency map to study the explainable AD evolution for all the patients in the study based on the predictions of individual biomarker models. BrainOS (2, 8, 14, 20) show the model prediction at different chunk sizes. This map could be used to longitudinally monitor subjects where a lower score on the biomarkers may indicate a more progressed AD subject. **(B)** Saliency map comparing AD+ subject S092 with a solid line and AD- subject S019 represented with a dashed line.

the sensory stream and cognitive modules are selected, we first concatenate them together and then pass their outputs through a 1,024 neuron deeply connected neural network layer with ReLU activation. We also incorporate at this point metadata such as gender. We test three Brain OS transfer learning strategies: (1) CNNs are used as fixed feature extractors without any fine-tuning; (2) CNNs are fine-tuned by training all layers; (3) only the final layers of the CNN are fine-tuned.

From **Figure 5**, it is evident AD detection improves as chunk length increases consistent with the fact that attention-marking has more per-chunk information to formulate a better AD prediction. From this attention-marking index (quantity of information required in a chunk for a confident diagnosis) we select chunk sizes **2**, **8**, **14**, and **20 s**, shown in **Figure 7**, as the Brain OS biomarkers, establishing individual AD progression. In terms of transfer learning strategies, **Figure 4** shows that fine-tuning all layers always leads to better results, however for most models almost no fine-tuning is required to beat the baseline.

6. OVBM COGNITIVE CORE BIOMARKERS

Neuropsychological tests are a common screening tool for AD (Baldwin and Farias, 2009). These tests, among others, evaluate a patient's ability to remember uncommon words, contextualize, infer actions, and detect saliency (Baldwin and Farias, 2009; Costa et al., 2017). In the case of this AD dataset, all patients are asked to describe the Cookie Theft picture created by Goodglass et al. (1983), where a set of words such as "kitchen" (**context**), "tipping" (**unique**), "jar" (**inferred**), and "overflow" (**salient**), are used to capture four cognitive biomarkers. To keep the richness of speech, we train four wake word models from LibriSpeech (Panayotov et al., 2015) with ResNet50s following the same approach as Biomarker 2. The four chosen cognitive biomarkers aim to detect patients' ability on: context, uniqueness, inference, and saliency.

We could show the same saliency bar chart in **Figure 4** and a uniqueness table such as **Table 3** to illustrate the impact of each cognitive biomarker. Instead in **Figure 5**, we show the impact of removing the cognitive core on the top OVBM performance which drops ≈10%, validating the relevance of the cognitive core biomarkers.

7. OVBM SYMBOLIC COMP. M. BIOMARKERS

This module fine-tunes previous modules' outputs into a graph neural network. Predictions on individual audio chunks for one subject are aggregated and fed into competing models to reach a final diagnostic. We tested the model with various BERT configurations and found no improvement in detection accuracy. In the AD implementation, given we had at most 39 overlapping chunks, three simple aggregation metrics are compared: averaging, linear positive (more weight given to later chunks), and linear negative (more weight given to earlier chunks).

In **Figure 6**, averaging proves to be the most effective, while positive linear over performing the negative linear indicates the latter audio chunks are more informative than front ones. **Figure 7** includes four biomarkers derived from combining chunk predictions from biomarker models of the three other modules (Cummings, 2019). With more data and longitudinal recordings, the OVBM GNN may incorporate other biomarkers.

8. DISCUSSION

We conclude by providing a few insights further supporting our OVBM brain-inspired model for audio health diagnostics as presented above. We have proven the success of the OVBM framework, setting the new benchmark for state-of-the-art accuracy of AD classification, despite only incorporating audio signals—one that can incorporate GNNs (Wu et al., 2020). Future work may improve this benchmark by also incorporating into OVBM longitudinal GNN's natural language biomarkers using NLP classifiers or multi-modal graph neural networks incorporating non-audio diagnostic tools (Parisot et al., 2018).

One of the most promising insights of all is the discovery of cough as a new biomarker (**Figure 3**), one that improves any of the intermediate models tested and that validates OVBM as a framework on which medical experts can hypothesize and test out existing and novel biomarkers. We are the first to report that cough biomarkers have information related to gender and culture, and are also the first to demonstrate how they improve simultaneous AD classification as illustrated in the saliency charts (**Figure 4**) as well as that of other apparently unrelated conditions.

Another promising finding is the model's explainability, introducing the biomarker AD saliency map tool, offering novel methods to evaluate patients longitudinally on a set of physical and neuropsychological biomarkers as shown in **Figure 7**. In future research, longitudinal data may be collected to properly test the onset potential of OVBM GNN discrimination in continuous speech. We hope our approach brings the AI health diagnostic experts closer to the medical community and accelerates research for treatments by providing longitudinal and explainable tracking metrics that can help succeed adaptive clinical trials of urgently needed innovative interventions.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: in order to gain access to the datasets used in the paper, researchers must become a member of DementiaBank. Requests to access these datasets should be directed to https://dementia.talkbank.org/.

AUTHOR CONTRIBUTIONS

JL wrote the code. BS designed the longitudinal biomarker Open Voice Brian Model (OVBM) and the saliency map. All authors contributed to the analysis of the results and the article.

REFERENCES

Abeyratne, U. R., Swarnkar, V., Setyati, A., and Triasih, R. (2013). Cough sound analysis can rapidly diagnose childhood pneumonia. *Ann. Biomed. Eng.* 41, 2448–2462. doi: 10.1007/s10439-013-0836-0

Alagiakrishnan, K., Bhanji, R. A., and Kurian, M. (2013). Evaluation and management of oropharyngeal dysphagia in different types of dementia: a systematic review. *Arch. Gerontol. Geriatr.* 56, 1–9. doi: 10.1016/j.archger.2012.04.011

Altinkaya, E., Polat, K., and Barakli, B. (2020). Detection of Alzheimer's disease and dementia states based on deep learning from MRI images: a comprehensive review. *J. Instit. Electron. Comput.* 1, 39–53. doi: 10.33969/JIEC.2019.11005

Alzheimer's Association (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's Dement.* 15, 321–387. doi: 10.1016/j.jalz.2019.01.010

Alzheimer's Association (2020). Alzheimer's disease facts and figures [published online ahead of print, 2020 mar 10]. *Alzheimers Dement.* 1–70. doi: 10.1002/alz.12068

Azarpazhooh, M. R., Amiri, A., Morovatdar, N., Steinwender, S., Ardani, A. R., Yassi, N., et al. (2020). Correlations between covid-19 and burden of dementia: an ecological study and review of literature. *J. Neurol. Sci.* 416:117013. doi: 10.1016/j.jns.2020.117013

Babulal, G. M., Ghoshal, N., Head, D., Vernon, E. K., Holtzman, D. M., Benzinger, T. L., et al. (2016). Mood changes in cognitively normal older adults are linked to Alzheimer disease biomarker levels. *Am. J. Geriatr. Psychiatry* 24, 1095–1104. doi: 10.1016/j.jagp.2016.04.004

Baldwin, S., and Farias, S. T. (2009). Unit 10.3: Assessment of cognitive impairments in the diagnosis of Alzheimer's disease. *Curr. Protoc. Neurosci.* 10:Unit10-3. doi: 10.1002/0471142301.ns1003s49

Barthélemy, N. R., Horie, K., Sato, C., and Bateman, R. J. (2020). Blood plasma phosphorylated-tau isoforms track CNS change in Alzheimer's disease. *J. Exp. Med.* 217:e20200861. doi: 10.1084/jem.20200861

Bennett, W. D., Daviskas, E., Hasani, A., Mortensen, J., Fleming, J., and Scheuch, G. (2010). Mucociliary and cough clearance as a biomarker for therapeutic development. *J. Aerosol Med. Pulmon. Drug Deliv.* 23, 261–272. doi: 10.1089/jamp.2010.0823

Bidzan, M., and Bidzan, L. (2014). Neurobehavioral manifestation in early period of Alzheimer disease and vascular dementia. *Psychiatr. Polska* 48, 319–330.

Bowling, A., Rowe, G., Adams, S., Sands, P., Samsi, K., Crane, M., et al. (2015). Quality of life in dementia: a systematically conducted narrative review of dementia-specific measurement scales. *Aging Ment. Health* 19, 13–31. doi: 10.1080/13607863.2014.915923

Briggs, R., Kennelly, S. P., and O'Neill, D. (2016). Drug treatments in Alzheimer's disease. *Clin. Med.* 16:247. doi: 10.7861/clinmedicine.16-3-247

Cano-Cordoba, F., Sarma, S., and Subirana, B. (2017). *Theory of Intelligence With Forgetting: Mathematical Theorems Explaining Human Universal Forgetting Using "Forgetting Neural Networks"*. Technical Report 71, MIT Center for Brains, Minds and Machines (CBMM).

Cassani, R., Estarellas, M., San-Martin, R., Fraga, F. J., and Falk, T. H. (2018). Systematic review on resting-state EEG for Alzheimer's disease diagnosis and progression assessment. *Dis. Mark.* 2018:5174815. doi: 10.1155/2018/5174815

Cera, M. L., Ortiz, K. Z., Bertolucci, P. H. F., and Minett, T. S. C. (2013). Speech and orofacial apraxias in Alzheimer's disease. *Int. Psychogeriatr.* 25, 1679–1685. doi: 10.1017/S1041610213000781

Chertkow, H., and Bub, D. (1990). Semantic memory loss in dementia of Alzheimer's type: what do various measures measure? *Brain* 113, 397–417. doi: 10.1093/brain/113.2.397

Clark, C. M., and Ewbank, D. C. (1996). Performance of the dementia severity rating scale: a caregiver questionnaire for rating severity in Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* 10, 31–39. doi: 10.1097/00002093-199603000-00006

Coffey, C. S., and Kairalla, J. A. (2008). Adaptive clinical trials. *Drugs R & D* 9, 229–242. doi: 10.2165/00126839-200809040-00003

Costa, A., Bak, T., Caffarra, P., Caltagirone, C., Ceccaldi, M., Collette, F., et al. (2017). The need for harmonisation and innovation of neuropsychological assessment in neurodegenerative dementias in Europe: consensus document of the joint program for neurodegenerative diseases working group. *Alzheimer's Res. Ther.* 9:27. doi: 10.1186/s13195-017-0254-x

Cummings, L. (2019). Describing the cookie theft picture: Sources of breakdown in Alzheimer's dementia. *Pragmat. Soc.* 10, 153–176. doi: 10.1075/ps.17011.cum

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J. Alzheimer's Dis.* 78, 1547–1574. doi: 10.3233/JAD-200888

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., et al. (2019). A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG pet of the brain. *Radiology* 290, 456–464. doi: 10.1148/radiol.2018180958

Dodd, J. W. (2015). Lung disease as a determinant of cognitive decline and dementia. *Alzheimer's Res. Ther.* 7:32. doi: 10.1186/s13195-015-0116-3

Ebihara, T., Gui, P., Ooyama, C., Kozaki, K., and Ebihara, S. (2020). Cough reflex sensitivity and urge-to-cough deterioration in dementia with Lewy bodies. *ERJ Open Res.* 6, 108–2019. doi: 10.1183/23120541.00108-2019

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Fuller, S. J., Carrigan, N., Sohrabi, H. R., and Martins, R. N. (2019). "Current and developing methods for diagnosing Alzheimer's disease," in *Neurodegeneration and Alzheimer's Disease: The Role of Diabetes, Genetics, Hormones, and Lifestyle,* eds R. N. Martins, C. S. Brennan, W. M. A. D. Binosha Fernando, M. A. Brennan, S. J. Fuller (John Wiley & Sons Ltd.), 43–87. doi: 10.1002/9781119356752.ch3

Galvin, J., Tariot, P., Parker, M. W., and Jicha, G. (2020). *Screen and Intervene: The Importance of Early Detection and Treatment of Alzheimer's Disease*. The Medical Roundtable General Medicine Edition.

Ghoniem, R. M. (2019). "Deep genetic algorithm-based voice pathology diagnostic system," in *Natural Language Processing and Information Systems*, eds E. Métais, F. Meziane, S. Vadera, V. Sugumaran, and M. Saraee (Cham: Springer International Publishing), 220–233. doi: 10.1007/978-3-030-23281-8_18

Giannakopoulos, P., Duc, M., Gold, G., Hof, P. R., Michel, J.-P., and Bouras, C. (1998). Pathologic correlates of apraxia in Alzheimer disease. *Arch. Neurol.* 55, 689–695. doi: 10.1001/archneur.55.5.689

Goodglass, H., Kaplan, E., and Barressi, B. (1983). *Cookie Theft Picture. Boston Diagnostic Aphasia Examination*. Philadelphia, PA: Lea & Febiger.

Guinjoan, S. M. (2021). Expert opinion in Alzheimer disease: the silent scream of patients and their family during coronavirus disease 2019 (covid-19) pandemic. *Pers. Med. Psychiatry* 2021:100071. doi: 10.1016/j.pmip.2021.100071

Hariyanto, T. I., Putri, C., Situmeang, R. F. V., and Kurniawan, A. (2020). Dementia is a predictor for mortality outcome from coronavirus disease 2019 (COVID-19) infection. *Eur. Arch. Psychiatry Clin. Neurosci.* 26, 1–3. doi: 10.1007/s00406-020-01205-z

Haulcy, R., and Glass, J. (2021). Classifying Alzheimer's disease using audio and text-based representations of speech. *Front. Psychol.* 11:3833. doi: 10.3389/fpsyg.2020.624137

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Heckman, P. R., Blokland, A., and Prickaerts, J. (2017). "From age-related cognitive decline to Alzheimer's disease: a translational overview of the potential role for phosphodiesterases," in *Phosphodiesterases: CNS Functions and Diseases* eds H. T. Zhang, Y. Xu, J. O'Donnell (Springer), 135–168. doi: 10.1007/978-3-319-58811-7_6

Henriksen, K., O'Bryant, S., Hampel, H., Trojanowski, J., Montine, T., Jeromin, A., et al. (2014). The future of blood-based biomarkers for Alzheimer's disease. *Alzheimer's Dement.* 10, 115–131. doi: 10.1016/j.jalz.2013.01.013

Holmes, R., Oates, J., Phyland, D., and Hughes, A. (2000). Voice characteristics in the progression of Parkinson's disease. *Int. J. Lang. Commun. Disord.* 35, 407–418. doi: 10.1080/136828200410654

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisc. Rev. Data Mining Knowl. Discov.* 9:e1312. doi: 10.1002/widm.1312

Hughes, C. P., Berg, L., Danziger, W., Coben, L. A., and Martin, R. L. (1982). A new clinical scale for the staging of dementia. *Br. J. Psychiatry* 140, 566–572. doi: 10.1192/bjp.140.6.566

Isaia, G., Marinello, R., Tibaldi, V., Tamone, C., and Bo, M. (2020). Atypical presentation of covid-19 in an older adult with severe Alzheimer disease. *Am. J. Geriatr. Psychiatry* 28, 790–791. doi: 10.1016/j.jagp.2020.04.018

James, H. J., Van Houtven, C. H., Lippmann, S., Burke, J. R., Shepherd-Banigan, M., Belanger, E., et al. (2020). How accurately do patients and their care partners report results of amyloid-β pet scans for Alzheimer's disease assessment? *J. Alzheimer's Dis.* 74, 625–636. doi: 10.3233/JAD-190922

Johnen, A., and Bertoux, M. (2019). Psychological and cognitive markers of behavioral variant frontotemporal dementia-a clinical neuropsychologist's view on diagnostic criteria and beyond. *Front. Neurol.* 10:594. doi: 10.3389/fneur.2019.00594

Kalia, M. (2003). Dysphagia and aspiration pneumonia in patients with Alzheimer's disease. *Metabolism* 52, 36–38. doi: 10.1016/S0026-0495(03)00300-7

Karlekar, S., Niu, T., and Bansal, M. (2018). Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. *arXiv preprint arXiv:1804.06440*. doi: 10.18653/v1/N18-2110

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. Available online at: https://arxiv.org/abs/1412.6980

Krijnders, J., and t Holt, G. (2017). Tone-fit and MFCC scene classification compared to human recognition. *Energy* 400:500. Available online at: https://www.researchgate.net/publication/255823915_Tonefit_and_MFCC_Scene_Classification_compared_to_Human_Recognition

Kuo, C.-L., Pilling, L. C., Atkins, J. L., Kuchel, G. A., and Melzer, D. (2020). !‘*i?‘APOE!‘/i?*‘ e2 and aging-related outcomes in 379,000 UK biobank participants. *Aging* 12, 12222–12233. doi: 10.18632/aging.103405

Laguarta, J., Hueto, F., and Subirana, B. (2020). Covid-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open J. Eng. Med. Biol.* 1, 275–281. doi: 10.1109/OJEMB.2020.3026928

Lambon Ralph, M. A., Patterson, K., Graham, N., Dawson, K., and Hodges, J. R. (2003). Homogeneity and heterogeneity in mild cognitive impairment and Alzheimer's disease: a cross-sectional and longitudinal study of 55 cases. *Brain* 126, 2350–2362. doi: 10.1093/brain/awg236

Lee, H., Pham, P., Largman, Y., and Ng, A. Y. (2009). "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems 22*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Curran Associates, Inc.), 1096–1104.

Livingstone, S. R., and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north American english. *PLoS ONE* 13:e196391. doi: 10.1371/journal.pone.0196391

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). "Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge," in *Proceedings of INTERSPEECH 2020* (Shanghai). doi: 10.21437/Interspeech.2020-2571

Lyons, J., Wang, D. Y.-B., Shteingart, H., Mavrinac, E., Gaurkar, Y., Watcharawisetkul, W., et al. (2020). jameslyons/python_speech_features: release v0.6.1.

Lyu, G. (2018). "A review of Alzheimer's disease classification using neuropsychological data and machine learning," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* Beijing, 1–5. doi: 10.1109/CISP-BMEI.2018.8633126

Maclin, J. M. A., Wang, T., and Xiao, S. (2019). Biomarkers for the diagnosis of Alzheimer's disease, dementia lewy body, frontotemporal dementia and vascular dementia. *Gen. Psychiatry* 32:e100054. doi: 10.1136/gpsych-2019-100054

Manabe, T., Fujikura, Y., Mizukami, K., Akatsu, H., and Kudo, K. (2019). Pneumonia-associated death in patients with dementia: a systematic review and meta-analysis. *PLoS ONE* 14:e0213825. doi: 10.1371/journal.pone.0213825

Michel, A., Verin, E., Hansen, K., Chassagne, P., and Roca, F. (2020). Buccofacial apraxia, oropharyngeal dysphagia, and dementia severity in community-dwelling elderly patients. *J. Geriatr. Psychiatry Neurol.* 34, 150–155. doi: 10.1177/0891988720915519

Morris, R. G., and Kopelman, M. D. (1986). The memory deficits in Alzheimer-type dementia: a review. *Q. J. Exp. Psychol.* 38, 575–602. doi: 10.1080/14640748608401615

Morsy, A., and Trippier, P. (2019). Current and emerging pharmacological targets for the treatment of Alzheimer's disease. *J. Alzheimer's Dis.* 72, 1–33. doi: 10.3233/JAD-190744

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: a systematic review. *IEEE Access* 7, 19143–19165. doi: 10.1109/ACCESS.2019.2896880

Olde Rikkert, M. G., Tona, K. D., Janssen, L., Burns, A., Lobo, A., Robert, P., et al. (2011). Validity, reliability, and feasibility of clinical staging scales in dementia: a systematic review. *Am. J. Alzheimer's Dis. Other Dement.* 26, 357–365. doi: 10.1177/1533317511418954

Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Baltimore, MD), 78–87. doi: 10.3115/v1/W14-3210

Palmqvist, S., Janelidze, S., Quiroz, Y. T., Zetterberg, H., Lopera, F., Stomrud, E., et al. (2020). Discriminative accuracy of plasma phospho-tau217 for Alzheimer disease vs. other neurodegenerative disorders. *JAMA* 324, 772–781. doi: 10.1001/jama.2020.12134

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (South Brisbane, QLD), 5206–5210. doi: 10.1109/ICASSP.2015.7178964

Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., et al. (2018). Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med. image Anal.* 48, 117–130. doi: 10.1016/j.media.2018.06.001

Pawlowski, M., Joksch, V., Wiendl, H., Meuth, S. G., Duning, T., and Johnen, A. (2019). Apraxia screening predicts Alzheimer pathology in frontotemporal dementia. *J. Neurol. Neurosurg. Psychiatry* 90, 562–569. doi: 10.1136/jnnp-2018-318470

Pearl, J., and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic books.

Petti, U., Baker, S., and Korhonen, A. (2020). A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J. Am. Med. Inform. Assoc.* 27, 1784–1797. doi: 10.1093/jamia/ocaa174

Pramono, R. X. A., Imtiaz, S. A., and Rodriguez-Villegas, E. (2016). A cough-based algorithm for automatic diagnosis of pertussis. *PLoS ONE* 11:e162128. doi: 10.1371/journal.pone.0162128

Pulido, M. L. B., Hernández, J. B. A., Ballester, M. Á. F., González, C. M. T., Mekyska, J., and Smékal, Z. (2020). Alzheimer's disease and automatic speech analysis: a review. *Expert Syst. Appl.* 2020:113213. doi: 10.1016/j.eswa.2020.113213

Reed, W. J., and Hughes, B. D. (2002). From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Phys. Rev. E* 66:067103. doi: 10.1103/PhysRevE.66.067103

Sekizawa, K., Jia, Y. X., Ebihara, T., Hirose, Y., Hirayama, Y., and Sasaki, H. (1996). Role of substance p in cough. *Pulmon. Pharmacol.* 9, 323–328. doi: 10.1006/pulp.1996.0042

Severini, C., Petrella, C., and Calissano, P. (2016). Substance p and Alzheimer's disease: emerging novel roles. *Curr. Alzheimer Res.* 13, 964–972. doi: 10.2174/1567205013666160401114039

Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., et al. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann. Neurol.* 65, 403–413. doi: 10.1002/ana.21610

Small, B. J., Fratiglioni, L., Viitanen, M., Winblad, B., and Bäckman, L. (2000). The course of cognitive impairment in preclinical Alzheimer disease: three- and 6-year follow-up of a population-based sample. *Arch. Neurol.* 57, 839–844. doi: 10.1001/archneur.57.6.839

Song, X., Mitnitski, A., and Rockwood, K. (2011). Nontraditional risk factors combine to predict Alzheimer disease and dementia. *Neurology* 77, 227–234. doi: 10.1212/WNL.0b013e318225c6bc

Subirana, B. (2020). Call for a wake standard for artificial intelligence. *Commun. ACM* 63, 32–35. doi: 10.1145/3402193

Subirana, B., Bagiati, A., and Sarma, S. (2017a). *On the Forgetting of College Academics: at "Ebbinghaus Speed"*? Technical Report 68, MIT Center for Brains, Minds and Machines (CBMM). doi: 10.21125/edulearn.2017.0672

Subirana, B., Bivings, R., and Sarma, S. (2020a). "Wake neutrality of artificial intelligence devices," in *Algorithms and Law*, eds M. Ebers and S. Navas (Cambridge University Press). doi: 10.1017/9781108347846.010

Subirana, B., Hueto, F., Rajasekaran, P., Laguarta, J., Puig, S., Malvehy, J., et al. (2020b). Hi Sigma, do I have the coronavirus?: call for a new artificial intelligence approach to support health care professionals dealing with the COVID-19 pandemic. *arXiv preprint arXiv:2004.06510*.

Subirana, B., Sarma, S., Cantwell, R., Stine, J., Taylor, M., Jacobs, K., et al. (2017b). *Time to Talk: The Future for Brands is Conversational*. Technical report, MIT Auto-ID Laboratory and Cap Gemini.

Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). "Automated screening for Alzheimer's dementia through spontaneous Speech," in *INTERSPEECH 2020 Conference* (Shanghai), 2222–2226. doi: 10.21437/Interspeech.2020-3158

The Center for Brains, Minds & Machines (2020). *Modules*. Available online at: https://cbmm.mit.edu/research/modules (accessed April 14, 2020).

Wirths, O., and Bayer, T. (2008). Motor impairment in Alzheimer's disease and transgenic Alzheimer's disease mouse models. *Genes Brain Behav*. 7, 1–5. doi: 10.1111/j.1601-183X.2007.00373.x

Wise, J. (2016). Dementia and flu are blamed for increase in deaths in 2015 in England and wales. *BMJ* 353:i2022. doi: 10.1136/bmj.i2022

Won, H.-K., Yoon, S.-J., and Song, W.-J. (2018). The double-sidedness of cough in the elderly. *Respir. Physiol. Neurobiol*. 257, 65–69. doi: 10.1016/j.resp.2018.01.009

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst*. 32, 4–24. doi: 10.1109/TNNLS.2020.2978386

Yu, Y., Travaglio, M., Popovic, R., Leal, N. S., and Martins, L. M. (2020). Alzheimer's and Parkinson's diseases predict different COVID-19 outcomes, a UK biobank study. *medRxiv*. 1–16. doi: 10.1101/2020.11.05.20226605

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease," in *INTERSPEECH 2020 Conference* (Shanghai), 2162–2166. doi: 10.21437/Interspeech.2020-2516

Zetterberg, H. (2019). Blood-based biomarkers for Alzheimer's disease–an update. *J. Neurosci. Methods* 319, 2–6. doi: 10.1016/j.jneumeth.2018.10.025

Zunic, A., Corcoran, P., and Spasic, I. (2020). Sentiment analysis in health and well-being: systematic review. *JMIR Med. Inform*. 8:e16023. doi: 10.2196/16023

Conflict of Interest: The 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.

ORIGINAL RESEARCH
published: 16 April 2021
doi: 10.3389/fcomp.2021.642517

Crossing the "Cookie Theft" Corpus Chasm: Applying What BERT Learns From Outside Data to the ADReSS Challenge Dementia Detection Task

Yue Guo[1]*, *Changye Li*[2], *Carol Roan*[3], *Serguei Pakhomov*[2] *and Trevor Cohen*[1]

[1] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States, [2] Pharmaceutical Care and Health Systems, University of Minnesota, Minneapolis, MN, United States, [3] Department of Sociology, University of Wisconsin-Madison, Madison, WI, United States

Large amounts of labeled data are a prerequisite to training accurate and reliable machine learning models. However, in the medical domain in particular, this is also a stumbling block as accurately labeled data are hard to obtain. DementiaBank, a publicly available corpus of spontaneous speech samples from a picture description task widely used to study Alzheimer's disease (AD) patients' language characteristics and for training classification models to distinguish patients with AD from healthy controls, is relatively small—a limitation that is further exacerbated when restricting to the balanced subset used in the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge. We build on previous work showing that the performance of traditional machine learning models on DementiaBank can be improved by the addition of normative data from other sources, evaluating the utility of such extrinsic data to further improve the performance of state-of-the-art deep learning based methods on the ADReSS challenge dementia detection task. To this end, we developed a new corpus of professionally transcribed recordings from the Wisconsin Longitudinal Study (WLS), resulting in 1366 additional Cookie Theft Task transcripts, increasing the available training data by an order of magnitude. Using these data in conjunction with DementiaBank is challenging because the WLS metadata corresponding to these transcripts do not contain dementia diagnoses. However, cognitive status of WLS participants can be inferred from results of several cognitive tests including semantic verbal fluency available in WLS data. In this work, we evaluate the utility of using the WLS 'controls' (participants without indications of abnormal cognitive status), and these data in conjunction with inferred 'cases' (participants with such indications) for training deep learning models to discriminate between language produced by patients with dementia and healthy controls. We find that incorporating WLS data during training a BERT model on ADReSS data improves its performance on the ADReSS dementia detection task, supporting the hypothesis that incorporating WLS data adds value in this context. We also demonstrate that weighted cost functions and additional prediction targets may be effective ways to address issues arising from class imbalance and confounding effects due to data provenance.

Keywords: dementia diagnosis, Alzheimer's disease, natural language processing, BERT, machine learning

OPEN ACCESS

Edited by:
Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:
Maria Koutsombogera,
Trinity College Dublin, Ireland
Heidi Christensen,
The University of Sheffield,
United Kingdom

****Correspondence:***
Yue Guo
yguo50@uw.edu

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *16 December 2020*
Accepted: *23 March 2021*
Published: *16 April 2021*

Citation:
Guo Y, Li C, Roan C, Pakhomov S and Cohen T (2021) Crossing the "Cookie Theft" Corpus Chasm: Applying What BERT Learns From Outside Data to the ADReSS Challenge Dementia Detection Task.
Front. Comput. Sci. 3:642517.
doi: 10.3389/fcomp.2021.642517

1. INTRODUCTION

Alzheimer's Dementia (AD) is a debilitating condition with few symptomatic treatments and no known cure. According to the Alzheimer's Association, in 2018 an estimated 5.8 million Americans were living with AD (Association, 2019). By 2050, these numbers are projected to increase to 14 million people with AD at a cost of $1.1 trillion per year (Association, 2019). Diagnosis of this condition is often missed or delayed (Bradford et al., 2009), and delays may occur over an extended period with cognitive changes anticipating future dementia preceding clinical diagnosis by as many as 18 years (Rajan et al., 2015; Aguirre-Acevedo et al., 2016). Earlier diagnosis of AD has the potential to ease the burden of disease on patients and caregivers by reducing family conflict and providing more time for financial and care planning (Boise et al., 1999; Bond et al., 2005; Stokes et al., 2015). Delayed diagnosis of this condition also contributes substantively to the cost of care of this disease on account of a high utilization of emergency rather than routine care, amongst other factors—it is estimated that early and accurate diagnosis can help save an estimated $7.9 trillion in medical and care costs (Association, 2018). Furthermore, survey findings show the vast majority (∼80%) would prefer to know if their unexplained symptoms of confusion or memory loss were due to AD dementia in a formal clinical evaluation (Blendon et al., 2011).

One path to earlier diagnosis of AD involves the application of machine learning methods to transcribed speech, with the publicly available DementiaBank corpus (Becker et al., 1994) providing a focal point for research in this area. The majority of this prior work has involved the application of supervised machine learning methods (see e.g., Orimaye et al., 2014, 2017, 2018; Fraser et al., 2016; Yancheva and Rudzicz, 2016; Karlekar et al., 2018; Cohen and Pakhomov, 2020) to classify groups of transcripts, specific transcripts or even individual utterances as to whether or not the participants producing them were clinically diagnosed with dementia. While many of the methods developed during the course of this research exhibited promising performance, their performance is not strictly comparable on account of differences in units of analysis, restrictions on the inclusion of participants, evaluation metrics and cross-validation strategies. Furthermore, the DementiaBank dataset was constructed without case/control matching, resulting in statistically significant differences in age and level of education across the AD and control groups. Consequently there is a danger that diagnostic performance of classifiers trained and evaluated on this set may be overestimated on account of their ability to learn to recognize these differences, rather than linguistic indicators of AD.

2. BACKGROUND

The ADReSS challenge reference set was deliberately constructed to remediate some of these issues with the original data (Luz et al., 2020). This dataset represents a subset of the DementiaBank data, matched for age and gender, with enhancement of the accompanying audio data, and containing only a single transcript for each participant (as opposed to the multiple transcripts corresponding to multiple study visits per participant available in the original set). As has been noted by the developers of the ADReSS dataset, it has the potential to advance the field by providing a standardized set for comparison between methods, which is a welcome advance on account of previously published work in this area often using different subsets of DementiaBank, as well as different cross-validation strategies and performance metrics. The ADReSS set and the accompanying challenge task present a standardized approach to evaluation on two tasks—AD recognition and Mini-mental State Exam (MMSE) prediction—for comparative evaluation moving forward. However, it is also true that this subset is even smaller in size than the original DementiaBank set, with only 108 training examples and 54 test examples, both split equally between healthy controls and participants with AD dementia.

In previous work, Noorian et al. (2017) demonstrated that the performance of machine learning approaches in the context of the DementiaBank set can be improved by providing the models concerned with additional "Cookie Theft" transcripts derived from other datasets. In this work, the authors introduced two additional sets of transcripts: Talk2Me and WLS. The former is an internal collection, while the latter is drawn from the Wisconsin Longitudinal Study (Herd et al., 2014), an extended study of a sample of students graduating from high school in Wisconsin 1957 born between 1938 and 1940 (initial n = 10,317), with some participants performing the "Cookie Theft" picture description task in a subsequent 2011 survey, aged in their early seventies. The authors report the availability of an additional 305 and 1,366 transcripts from participants without AD in the Talk2Me and WLS sets, respectively. In both cases, only recordings were available for analysis—text features were extracted using the Kaldi open source Automated Speech Recognition (ASR) engine (Povey et al., 2011), with an estimated word error rate of ∼12.5% on the Talk2Me data, and none provided for the WLS set. As the additional data were considered as controls, the ADASYN (He et al., 2008) synthetic sampling method was used to oversample the minority "dementia" class. On a random 80/20 train/test split of the DementiaBank data, the authors report a considerable advantage in performance with the addition of the WLS controls in particular, with improvements of over 10% (absolute) in macro-averaged F-measure across a range of machine learning methods trained on a set of 567 manually engineered features, with oversampling offering an advantage over training without balancing the set in some but not all methods.

In this paper we evaluate the extent to which the performance of contemporary deep learning architectures can benefit from the addition of data from the WLS set. After attaining the relevant institutional approvals, we obtained all available "Cookie Theft" recordings from the WLS collection, as well as professional transcriptions of these recordings, to obviate the need to consider ASR error in our subsequent analyses. We evaluate the utility of the resulting transcripts as a means to improve performance of transfer learning using pre-trained Transformer-based architectures (Vaswani et al., 2017), focusing on the widely-used Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018) that has been shown to outperform other machine learning methods on the ADReSS challenge in recent work (Balagopalan et al., 2020).

Combining text corpora drawn from different sources to train NLP models should be approached with caution. Recent NLP research has identified and attempted to address the potentially deleterious role of confounding variables in text classification (Landeiro and Culotta, 2018). A confounding variable is a variable that can influence both a predictor and an outcome of a predictive model. One manifestation of the issue of confounding in NLP concerns a scenario in which data are drawn from different sources (Howell et al., 2020), each with different underlying class distributions. The WLS and ADReSS sets exemplify this problem. The ADReSS set is balanced by design, with an equal number of case and control transcripts. However, while some indication of cognitive impairment can be inferred from the metadata that accompanied the WLS recordings, the control transcripts vastly outnumber the cases in which data from cognitive tasks indicates such impairment. Consequently, if differences in language use across the populations from which these datasets are drawn were to permit a machine learning model to distinguish between the two sets, such a model may approach its optimization objective of accurate classification by simply learning to label all WLS examples as controls. In this context, the provenance of a transcript serves as a confounding variable, because it influences both the intended predictors (words in the transcript) and the outcome of interest (whether or not the transcript was produced by a healthy control). In the context of deep neural networks for image recognition, it has been proposed that the problem of confounding can be addressed by introducing confounding variables of interest as additional model outputs (Zhong and Ettinger, 2017). The authors of this work argue that including confounding variables as secondary prediction objectives will influence model weights via backpropagation, resulting in models with better generalizability and overall performance. This argument is supported by empirical results demonstrating improved performance on an image classification task when potential confounding variables indicating position and orientation are incorporated as secondary targets for prediction. Motivated by this argument, we evaluate the utility of treating the provenance of a transcript (ADReSS vs. WLS) as a secondary target for prediction on overall model performance, with the secondary objective of determining the extent to which deep neural networks can learn to distinguish between unseen transcripts from each of these corpora. This secondary objective is of interest because accurate classification of unseen transcripts would confirm that there is systematic difference between transcripts from each corpus that has the potential to bias machine learning models, despite this not being immediately apparent upon qualitative evaluation of randomly selected transcripts during the process of data preparation.

A second concern with combining transcripts in this manner is that it introduces a class imbalance, where transcripts from healthy "controls" greatly outnumber those from patients with dementia. Previous work with WLS data used oversampling of the minority class to address this imbalance, which was effective with some but not all models (Noorian et al., 2017). As recent work with BERT suggests cost-sensitive learning is an effective alternative to address class imbalance (Madabushi et al., 2019), we evaluate the utility of this method also. Cost-sensitive learning involves adjusting the loss function of a model such that changes in performance on one class are weighted more heavily. In this case this involves proportionally weighting the loss function as an inverse function of the class distribution, such that the model learns to avoid misclassifying transcripts from dementia patients more assiduously than it learns not to misclassify those from healthy controls. Finally, we note that unlike the ADReSS set, the WLS transcripts do not come with diagnostic labels. However, the metadata accompanying these transcripts do include results of verbal fluency tests, as well as metadata indicative of clinical diagnoses other than dementia. A straightforward way to use these metadata involves developing an exclusion criterion, such that transcripts from participants with verbal fluency scores suggestive of diminished cognitive function are not treated as controls. In an additional effort to address the class imbalance introduced by the WLS data, we also experiment with treating the below-threshold fluency scores appended to these excluded transcripts as "noisy labels" (Natarajan et al., 2013) for the presence of dementia.

Thus, our research aims to answer the following key questions:

1. Does the performance of contemporary deep learning models on the ADReSS challenge diagnosis task benefit from the introduction of additional normative data comprising of "Cookie Theft" recordings from outside the ADReSS (or DementiaBank) set?
2. Does the addition of auxiliary outputs, or the incorporation of a cost-sensitive weight function, provide a way to compensate for the potential confounding effects and class imbalance introduced by these additional normative data, respectively?
3. Can verbal fluency scores be used to derive "noisy labels" to produce additional "case" training examples that are of value for performance on this task?
4. Are the two corpora sufficiently different that a deep learning model might learn to distinguish between them, during the course of the classification procedure?

Our main contributions can be summarized as follows:

1. We introduce a new professionally transcribed data set of 1,366 transcripts of the "Cookie Theft" task
2. We use associated metadata to infer noisy "case" and "control" labels for each transcript
3. We evaluate the utility of these additional data with and without inferred labels to improve the performance of transfer learning approaches on the ADReSS challenge classification task
4. We compare a set of loss function alternatives as a means to further improve performance.

3. MATERIALS AND METHODS

3.1. Dataset

3.1.1. ADReSS

The ADReSS dataset, derived from the DementiaBank dataset, consists of 156 speech transcriptions from AD and non-AD patients which are matched for age and gender. Transcripts are English language responses to the "Cookie Theft" task of the Boston Diagnostic Aphasia Exam, and are classified as "AD" or "control" on the basis of clinical and/or pathological examination. We downloaded the ADReSS dataset from the *Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge* website[1].

3.1.2. Wisconsin Longitudinal Study

The Wisconsin Longitudinal Study (WLS) is a longitudinal study of a random sample of 10,317 graduates from Wisconsin high schools in 1957. The study also includes a randomly selected sibling of graduates, and spouses of graduates and siblings. WLS participants were interviewed up to six times across 60-years between 1957 and 2011. Beginning in 1993, during the fourth round of interviews, the WLS included cognitive evaluations. The "Cookie Theft" task was administered in the sixth-round of the survey in 2011 survey (see Herd et al., 2014 for details). In July of 2019 the ongoing seventh round of data collection began.

3.2. Experiments

3.2.1. Dataset Construction

All audio samples in the WLS dataset were transcribed near-verbatim by a professional service. The resulting near-verbatim transcripts include filled pauses (um's and ah's) and tags for unintelligible speech. The transcriptionists also separated the speech of the examiner (containing task instructions and task-final comments) from the participant's response to the task. For the purposes of the current study, we removed filled pauses and unintelligible speech segments as well as the text corresponding to the examiner's speech.

The metadata of WLS do not currently provide dementia-related diagnoses; however, they do provide a limited set of cognitive test scores, and answers to questions about some health conditions. Of relevance to the current research, WLS participants underwent two category verbal fluency cognitive tests in which they were asked to name all words that belonged to a category (animals, food) in 1 min. The semantic (category) verbal fluency task has been previously shown to be highly sensitive (albeit not specific) to manifestations of AD dementia (Henry et al., 2004) with an unadjusted for age and education cutoff of 15 on the animal category recommended for use as a screening instrument in a clinical setting (Duff-Canning et al., 2004).

In order to identify a subgroup of healthy controls in the WLS dataset comparable to controls in the ADReSS dataset we used the verbal fluency scores and an answer of "yes" to the question "Have you ever been diagnosed with mental illness?" as inclusion/exclusion criteria as follows. We classified transcripts of participants as cases (as opposed to healthy controls) if (1) the participants had evidence of impairment in semantic verbal fluency, or (2) have been diagnosed with a mental illness[2]. Prior work on verbal fluency performance in participants with AD established that animal fluency scores <15 are 20 times more likely in a patient with AD than in an healthy individual and were found to discriminate between these two groups with sensitivity of 0.88 and specificity of 0.96 (Duff-Canning et al., 2004). Recognizing the fact that verbal fluency performance does vary slightly by age and education (Tombaugh et al., 1999; Marceaux et al., 2019), we used statistically determined age and education-adjusted thresholds of 16, 14, and 12 for participants in <60, 60–79, and >79 age ranges, respectively. We did not have normative data available for the food category; however, since the distributions of semantic verbal fluency scores on the "animal" category and "food" category were very similar, we applied the same cutoffs for the food category as for the animal category.

The initial set of 1,366 WLS participants was reduced to 1,165 by removing those with extremely long and short transcripts whose length was beyond one standard deviation around the mean length of a WLS transcript. Of the remaining WLS participants with a "Cookie Theft" picture description task transcription, 954 participants also had a category semantic verbal fluency score or indicated a mental illness diagnosis. Of these participants, 839 had a verbal fluency score above the normative threshold and did not have a mental illness diagnosis. These were labeled as "controls." Of the remaining 115 participants, 98 had a verbal fluency score below the threshold and 20 had a mental illness diagnosis. These 115 participants were labeled as "cases."

Descriptive statistics for the ADReSS and WLS datasets are shown in **Table 1**. The mean ages of WLS controls and cases at the point of data collection are lower than those of participants whose transcripts make up the ADReSS dataset. Upon analysis of the differences in age of participants between the two corpora, we found that while there was no statistically significant difference $[t(1108) = 4.3, p = 1.96]$ in the overall age of ADReSS ($M = 65.6$, $SD = 6.6$) and WLS participants ($M = 63.9$, $SD = 4.1$), nor in the age of controls $[t(915) = 1.3, p = 0.19]$[3], there was a significant difference between the ages of AD cases in the ADReSS set and inferred WLS "cases" $[t(191) = 4.6, p < 0.001]$. While statistically significant, this difference in mean ages is relatively small (2.3 years) and may be of limited practical significance. Gender distributions among these two datasets are similar. In both the WLS and ADReSS sets, a larger proportion of the control group attained post-high-school education.

3.2.2. Model

Bidirectional Encoder Representations from Transformers (*BERT* Devlin et al., 2018) provides a pretrained deep neural network for researchers and practitioners to fine tune on specific tasks by adding just one additional output layer (Liu and Lapata, 2019). BERT exemplifies the "transfer learning" approach that has been used to improve performance across a range

[1] http://www.homepages.ed.ac.uk/sluzfil/ADReSS/

[2] We use a generic term "cases" for participants with potential cognitive impairment and mental illness only as a way to distinguish them from controls, as we expect their language production on the picture description task to differ from that of controls. We do not in any way imply that a mental illness diagnosis is related to cognitive impairment. However, in the absence of specific metadata related to the presence of dementia, we decided it would be better to exclude these participants from the control set also.

[3] T-test results are reported in APA style: t(degrees of freedom) = the t statistic, p = p-value. The abbreviations M and SD stand for mean and standard deviation, respectively.

TABLE 1 | Dataset description.

		ADReSS			WLS		
		Control	Case	*P*-value	Control	Case	*P*-value
n		**78**	**78**		**839**	**115**	
Age, mean (SD)		65.0 (7)	66.3 (7)	0.226	63.9 (5)	63.9 (4)	0.902
Gender, *n* (%)	Female	43 (55)	43 (55)	1	295 (35)	32 (28)	0.995
	Male	35 (45)	35 (45)		213 (25)	23 (20)	
	Refused/Missing	0	0		331 (40)	60(52)	
Education, *n* (%)	≤12 years	34 (44)	52 (67)	0.002	401 (48)	78 (68)	<0.001
	>12 years	43 (55)	22 (28)		438 (52)	37 (32)	
	Refused/Missing	1 (1)	4 (5)		0	0	

Two-sample t*-tests were used to evaluate the* p*-value for continuous variables, and Chi-squared was used for categorical variables.*

of classification tasks in image and text processing in recent years. Essentially, transfer learning allows for the application of information learned while training a model on one task, to a different one. In the case of BERT for text classification, the initial task involves predicting held out ("masked") words or sentences in a large corpus of otherwise unlabeled text. The general information about word distribution and relative position learned in this manner can then be applied to a downstream classification task, with or without fine-tuning the weights of BERT in addition to a classification layer that is appended to this pretrained deep neural network model. Unlike previous recurrent neural network approaches, BERT allows the model to process words in relation to all other words in a passage in parallel rather than sequentially, enhancing the scalability of the pre-training procedure. An important feature of BERT is its use of attention modules (Vaswani et al., 2017), which take into account other words in a unit of text when generating a word representation during pre-training and subsequent tasks. BERT can therefore take the broader context of a word into consideration, with the capacity to resolve ambiguities in contextual word meaning. Most importantly, the information acquired during the pre-training process enables BERT to perform well even when only small amounts of annotated data are available for fine tuning. Following previous work, we modified BERT by adding a classification layer, to obtain binary class labels corresponding to "cases" and "controls" in the ADReSS dataset.

3.2.3. Loss Functions

We evaluated the utility of several variants of the BERT loss function as a means to compensate for class imbalance, and potential confounding effects. The standard loss function for categorization with BERT (as implemented in the widely used Hugginface Transformers library[4]) is the `CrossEntropy` loss, which combines a softmax function with the standard Cross Entropy loss. This encourages a model to choose one of a set of possible classes in a text categorization class, by converting model outputs into a series of probabilities across classes, which sum to one across all classes, before calculating the loss. For multi-label classification, where more than one label can be assigned (in our case, diagnosis = [case|control], source = [WLS|ADRess]), a reasonable alternative is to use the `BCEwithLogits` (BCE) loss function, which does not require probabilities as inputs. As this loss function also provides a convenient means to weight classes, we retained it for our experiments with cost-weighting as a means to compensate for class imbalance by applying a weight of $\frac{n}{c}$ for each class, where n is the number of transcripts in the set, and c is the number of transcripts of the class of interest. Less frequent classes (the "dementia" class when WLS is used) will have more influence on the cost function, as they will have a smaller denominator. In order to isolate the effects of this loss function from the multilabel and weighted configurations of it, we also report results with an unweighted edition of the `BCEwithLogits` loss, as well as the standard loss function.

3.2.4. Methods and Evaluation

To evaluate the effect of adding more data, the WLS control and WLS total sets (case and control) were added to the ADReSS training set separately. We used the single unique ADReSS test set as the testing set for all models, and evaluated the models by accuracy and area under the receiver-operator curve (AUC). We also performed cross-validation (CV) on the training set.

We report evaluation metrics with 5-fold CV (rather than the leave-one-subject-out protocol used in some prior work) due to memory and time constraints. In this case, values of evaluation metrics were averaged across CV folds. To evaluate performance on the test set, we generated 10 instantiations of each model using different random seeds to determine the initialization of classifier weights for each instantiation. We trained each of these models on the training set (± the WLS components) and reported the mean and standard error across these ten runs. For two class label prediction, we evaluated models with the standard loss function, a weighted BCE loss function, and an unweighted BCE loss function. Finally, we evaluated a multi-label classification model (AD, not AD, ADReSS, WLS), using an unweighted BCE loss function.

3.2.5. Training Details

All experiments were conducted with the 12-layer `bert-base-uncased` model. Experiments using cross-validation on the training set were run on a single NVIDIA Tesla

[4] https://github.com/huggingface/transformers

P-40 GPU, while experiments with evaluation on the test set were run on a single NVIDIA Tesla V-100 GPU. All models were developed using Python 3.7 and `PyTorch` 1.2.0. We used the `Transformers` library to implement BERT in PyTorch (Wolf et al., 2019), permitting fine-tuning of BERT model weights in addition to tuning of the classification layer. The maximum sentence length was set to the maximum length of the current training set, and the batch size was set to 8. The learning rate was set to 1×10^{-5}. All models were run for 20 epochs. We adopted the Adam optimizer (Kingma and Ba, 2014) with linear scheduling (Paszke et al., 2019) of the learning rate. For the BCE loss function, `nn.BCEWithLogitsLoss` was used. Other hyper-parameters were set to their default values.

4. RESULTS

The results of our 5-fold cross-validation experiments are shown in **Table 2**. When interpreting this table it is important to bear in mind that the cross-validation splits in the WLS control and WLS total scenarios include examples from the respective WLS sets also. Thus, they are not comparable to one another, nor are they comparable to the results shown with the ADReSS set only. However, it is informative to compare the results within each panel in turn (aside from the ADReSS-only result, which provides an indication of the robustness of the results from the train-test split used in the challenge). It is also important to note that the standard error of the mean (indicated with ±) is calculated across the five cross-validation folds, and consequently are indicative of the differences between the validation sets in these folds, rather than differences emerging from stochastic initialization of the classification layer of the BERT models concerned (these were initialized with the same random seed).

Both the WLS control and WLS total results suggest a trend toward an advantage for the loss function variants under consideration, as compared with the standard loss function, with unweighted and weighted variants of the BCE loss function generally outperforming the standard loss function. In addition, the best results in most cases are attained by the multilabel model. This suggests that augmentation of the model with additional targets for prediction may be helpful to reduce the confounding effect of the provenance of the transcripts concerned, when transcripts from both sources are included in the validation set. However, we note that one exception to this finding is the AUC in the WLS total set—in this configuration, the standard loss function performs best. The relatively poor performance with the addition of the "WLS total" set in 5-fold CV may result from discrepancies between the noisily labeled WLS cases and the clinically determined ADReSS AD dementia cases.

Results on the held-out ADReSS challenge test set are shown in **Table 3**, with the model trained on the ADReSS training set only and using the standard loss function taken as a baseline (these baseline results are largely consistent with the 5-fold cross-validation results on this set, suggesting the test set is representative of the data set as a whole). When comparing results from the three models trained with a standard loss function to evaluate the impact of the WLS data on a standard BERT classifier, we find both incorporating additional WLS controls, and the WLS total data (with controls and noisy labels for cases) leads to improvements over the baseline model. On account of the small number of test cases, only the advantages in AUC are statistically significant—presumably on account of accuracy generally having higher variance across runs than the AUC (as a small change in the predicted probability of an example may lead to a larger change in accuracy if this crosses the classification boundary and leads to error). Nonetheless, the general trend supports the hypothesis that the additional normative data will improve the performance of BERT on the ADReSS challenge diagnosis task.

When comparing the loss function variants, we observe that those models trained on the ADReSS set with the addition of WLS controls only using the weighted BCE function achieves the best AUC and accuracy amongst all the models, suggesting that weighting the loss function is an effective way of compensating for the class imbalance that results from these additional "control" data points. More importantly, this model significantly improves AUC compared to the baseline model in the test set. Unlike the 5-fold CV scenario, the multi-label loss function does

TABLE 2 | Five-fold cross-validation results on training set.

Data	Loss function	% Accuracy	% AUC
ADReSS	Standard	80.5 ± 4.0	88.2 ± 3.1
ADReSS + WLS control	Standard	96.5 ± 0.3	98.7 ± 0.3
	Weighted BCE	97.4 ± 0.3	98.9 ± 0.2
	Unweighted BCE	97.4 ± 0.3	98.8 ± 0.4
	Multilabel BCE	**97.9 ± 0.5**	**99.2 ± 0.1**
ADReSS + WLS total	Standard	83.3 ± 1.2	**68.8 ± 0.6**
	Weighted BCE	83.7 ± 1.4	61.2 ± 2.8
	Unweighted BCE	83.7 ± 1.4	65.7 ± 2.8
	Multilabel BCE	**84.8 ± 1.3**	66.1 ± 1.6

Results shown are the mean across the 5-folds ± the standard error. Best results in panels showing multiple models are in boldface.

TABLE 3 | Results on ADReSS test set.

Data	Loss function	% Accuracy	% AUC
ADReSS	Standard	79.8 ± 0.9	88.3 ± 0.5
ADReSS + WLS control	Standard	81.2 ± 1.1	90.6 ± 0.9
	Weighted BCE	**82.1 ± 1.0**	**92.3 ± 0.4***
	Unweighted BCE	80.8 ± 1.1	91.6 ± 0.3*
	Multilabel BCE	81.2 ± 0.5	90.6 ± 0.5*
ADReSS + WLS total	Standard	**81.9 ± 1.1**	**91.2 ± 0.9***
	Weighted BCE	80.8 ± 0.6	89.3 ± 0.9
	Unweighted BCE	80.8 ± 1.1	88.9 ± 0.5
	Multilabel BCE	80.4 ± 0.9	**91.2 ± 0.4***

*Results shown are the mean across ten iterations ± the standard error. *Indicates statistically significant difference from the baseline, as estimated by a paired t-test (as each repeated train/test evaluation was initialized with the same random seed across models). Best results in panels showing multiple models are in boldface.*

not lead to better performance than the standard loss function—which is perhaps unsurprising given the total absence of WLS data in the test set, obviating the need to resolve confounding effects emerging from data provenance. That the unweighted BCE function also does not improve performance over the standard function supports the hypotheses that it is indeed the weighting of this function that is responsible for its advantages in performance.

An additional finding from these experiments is that the multilabel models correctly identified the provenance of the ADReSS-derived examples in the test set with perfect accuracy in nine of 10 runs, and ∼98% accuracy on the remaining run. These results strongly support the hypothesis that a deep learning model trained on data from both corpora would learn to distinguish between them. This finding is further supported by a perfect accuracy in distinguishing between these corpora in the held-out validation split (including both WLS- and ADReSS-derived examples) demonstrated in a subsequent run.

The results with the inclusion of "noisy" WLS cases differ from those with controls alone. With the standard loss function, the addition of these data improves performance beyond that attained by adding WLS controls alone. However, performance does not match the best of the "control only" models, and is not improved further with the addition of variant loss functions. One explanation for the latter finding may be that class imbalance effects are already obviated through the introduction of additional cases, increasing the positively labeled training examples from ∼5 to ∼16% of the data available for training.

5. DISCUSSION

In this paper, we evaluated the utility of the incorporation of additional "Cookie Theft" transcripts drawn from the Wisconsin Longitudinal Study as a means to improve the performance of a BERT-based classifier on the ADReSS challenge diagnosis task. Our aims in doing so were primarily to evaluate whether or not these data would improve performance, but also to establish the extent to which weighting the cost function of the model and representing corpus provenance as additional targets for prediction could compensate for the issues of class imbalance and corpus-specific confounding effects, respectively. Finally, we wished to determine whether or not a model could learn to distinguish between the two corpora, to determine if our concern about such corpus-specific confounding effects was justified.

We found that incorporation of WLS data improved performance over that of a model trained on ADReSS data alone, and that these improvements were present both when only WLS "controls" (transcripts from participants with verbal fluency scores in the normal range for their age) were added, and when these were combined with noisily-labeled WLS "cases" (transcripts from participants with low verbal fluency scores, or reported diagnoses related to mental illness). When only controls were added, further improvements in performance were obtained when weighting the cost function to compensate for class imbalance, resulting in the best-performing models on the ADReSS challenge test set, with a mean accuracy of 82.1% and mean AUC of 92.3% across ten repeated instantiations of the model, as opposed to a mean accuracy and AUC of 79.8 and 88.3%, respectively, without the addition of WLS data.

While we note that Balagopalan et al. (2020) report an accuracy of 83.3% with a BERT-based classifier on this task when trained on the AD set alone, these experiments did not include repeated model instantiations to determine the effects of stochastic initialization of classifier weights on performance, that our baseline AD-only BERT model attained an accuracy of 83.3 or higher on two of ten such iterations, and that the best performance of the cost-weighted model across iterations resulted in an accuracy of 89.6% (with an AUC of 94.8%). This difference in baseline performance may be attributable to differences in stochastic initialization, or an unspecified difference in model architecture (e.g., BERT-base vs. BERT-large) or hyperparameter settings, and we do not believe it detracts from the strength of our conclusions.

While most of the work with the ADReSS challenge data has focused on multimodal analyses of acoustic and transcript data simultaneously, the paper introducing this data set provides some baseline results with language-only models, which were trained on a set of thirty-four linguistic outcome measures (such as total number of utterances, and part-of-speech percentage) (Luz et al., 2020). Test set classification accuracy is generally lower than results attained using BERT trained on raw text (even without the addition of WLS data), ranging from 0.625 to 0.792 across algorithms. Best performance was attained using a Support Vector Machine, though this configuration performed worse than other algorithms in cross-validation experiments. This suggests that BERT is able to automatically extract predictive features that outperform handcrafted features. However, it should be noted that BERT has considerably more trainable parameters than the models evaluated in this prior work, and that a fair comparison between BERT-based and engineered features would require the ascertainment of BERT's performance with freezing of all layers aside from the classification layer. Other work focusing on linguistic features explores the utility of using terms as features directly. Searle et al. (2020) compared machine learning models applied to word-level features to the DistilBERT architecture, reporting tied best accuracies of 81% with DistilBERT and an utterance-level combination of a Support Vector Machine and Conditional Random Field classifier. Additional work suggests that incorporation of acoustic features may offer further advantages in performance. Syed et al. (2020) demonstrated an accuracy of 85.4% with a multimodal learning system that incorporated both audio signals and transcripts. BERT and RoBERTa were included in the multimodal framework. These results suggest that incorporating additional information from auditory features may suggest a path toward further improving the performance of our models, although there are technical challenges concerning the differences in recording instruments across data sets that would need to be addressed in order to explore this.

In the context of 5-fold cross-validation experiments, where both WLS- and ADReSS-derived examples were present in the validation splits, adding transcript provenance as an additional target for prediction in a multi-label setting resulted in best performance, supporting the hypotheses that this may be an effective way to address corpus-specific confounding effects, which are an important concern in biomedical machine learning when there is a need to assemble a data set from smaller constituents that may have been collected at different institutions. Of particular interest for future work, these models also learned to classify the provenance of the data sets concerned with perfect or near-perfect accuracy, suggesting systematic differences between the source corpora that were not apparent upon informal inspection of word usage and lexical patterns. Further research is required to determine the cues used by the models to make these distinctions.

The results presented in this paper should be interpreted in light of several limitations. First, the ADReSS dataset is relatively small. The results reported here need to be replicated on larger datasets to determine their generalizability. Second, while the WLS dataset contains a very rich set of participant characteristics, these characteristics do not include those that can be used directly for characterization of AD dementia. Thus, the results pertaining to WLS cases should be interpreted with caution. In particular, while utilization of mental health diagnoses to exclude transcripts from the analysis is readily justifiable, using these diagnoses to derive noisy labels for cases may exceed the bounds of noise that our models can tolerate. In future work we will evaluate the extent to which using verbal fluency derived criterion only leads to noisy labels with greater downstream utility. We note also that efforts are currently underway to interview WLS participants in order to obtain clinical diagnoses of dementia. We anticipate this measure will be available for all eligible participants within a few years of the time of this writing, which will further enhance the utility of our transcripts for future research on the linguistic manifestations of dementia. Third, cases in both datasets are significantly less educated than the controls which may result in language use artifacts that have not been accounted for. These potential differences in language use should be further investigated. There are also some methodological alternatives that we did not fully consider in the current work. Our study did not consider the acoustic components of the available data, and depended upon manual transcriptions of speech data. Further research is needed to determine the utility of incorporating acoustic features, as well as the model's robustness to errors that may be introduced during the process of automated speech recognition. Furthermore, we did not formally evaluate oversampling strategies. Preliminary experiments with random oversampling suggested this would not be a fruitful strategy, and to our knowledge BERT-based strategies for similarity-based oversampling have yet to be developed. In addition we have yet to evaluate the combination of auxiliary prediction targets with weighting of the cost function, which may be a productive direction to pursue in future work on account of their individual utility when transcripts from both corpora are present at the point of validation. Finally, the utility of auxiliary targets as a means to obviate for confounding effects may be more readily apparent when the distribution of positive cases across corpora is different at test time than at training time (Landeiro and Culotta, 2018). Establishing whether or not this is the case would require additional evaluation involving validation sets in which these distributions are artificially modified.

6. CONCLUSION

In this paper, we evaluated the utility of using additional "Cookie Theft" picture description transcripts from the Wisconsin Longitudinal Study, as a means to improve the performance of a BERT-based classification approach on the dementia detection task of the ADReSS challenge. Our results indicate that training on these additional data leads to improved performance on this task, both when using all available transcripts as normative data regardless of cognitive status and subsets of the data extracted based on cognitive status inferred from available metadata (i.e., verbal fluency and mental health status). In the former case in particular, we find that weighted cost functions are an effective way to compensate for the class imbalance introduced by the addition of more "control" transcripts. Furthermore, results from our cross-validation studies suggest that introducing dataset provenance as an auxiliary target for prediction shows potential as a means to address different case/control distributions when combining datasets drawn from different sources. As such, our results suggest that our professionally transcribed WLS "Cookie Theft" transcripts are a valuable resource for the development of models to detect linguistic anomalies in dementia. These transcriptions are available upon request from wls@ssc.wisc.edu.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS), http://www.homepages.ed.ac.uk/sluzfil/ADReSS/. Requests to access our professional transcriptions of the Wisconsin Longitudinal Study (WLS) data should be directed to wls@ssc.wisc.edu.

AUTHOR CONTRIBUTIONS

YG, CL, TC, and SP conceived of the study design. YG and TC performed the experiments, with analysis of the results conducted in collaboration with authors SP and CL. CR provided the access to the WLS recordings and metadata, as well as guidance in the analysis and interpretation of these metadata. YG wrote the initial draft of the manuscript, with input from TC. All authors read and reviewed the manuscript, providing edits and suggestions for improvement where appropriate.

FUNDING

This research was supported by Administrative Supplement R01 LM011563 S1 from the National Library of Medicine, and R21 AG069792 from the National Institute of Aging. Since 1991, the WLS has been supported principally by the National Institute on Aging (AG-9775, AG-21079, and AG-033285), with additional support from the Vilas Estate Trust, the National Science Foundation, the Spencer Foundation, and the Graduate School of the University of Wisconsin-Madison.

REFERENCES

Aguirre-Acevedo, D. C., Lopera, F., Henao, E., Tirado, V., Muñoz, C., Giraldo, M., et al. (2016). Cognitive decline in a colombian kindred with autosomal dominant alzheimer disease: a retrospective cohort study. *JAMA Neurol.* 73, 431–438. doi: 10.1001/jamaneurol.2015.4851

Association, A. (2018). 2018 Alzheimer's disease facts and figures. *Alzheimers Dement.* 14, 367–429. doi: 10.1016/j.jalz.2018.02.001

Association, A. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers Dement.* 15, 321–387. doi: 10.1016/j.jalz.2019.01.010

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). To bert or not to bert: comparing speech and language-based approaches for Alzheimer's disease detection. *arXiv [Preprint] arXiv:2008.01551.* doi: 10.21437/Interspeech.2020-2557

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archiv. Neurol.* 51, 585–594. doi: 10.1001/archneur.1994.00540180063015

Blendon, R., Benson, J., Wikler, E., Weldon, K., Baumgart, M., Jansen, S., et al. (2011). Five-country survey of public experiences, attitudes and beliefs concerning Alzheimer's disease and the value of a diagnosis. *Alzheimers Dement.* 7:e50. doi: 10.1016/j.jalz.2011.09.209

Boise, L., Camicioli, R., Morgan, D. L., Rose, J. H., and Congleton, L. (1999). Diagnosing dementia: perspectives of primary care physicians. *Gerontologist* 39, 457–464. doi: 10.1093/geront/39.4.457

Bond, J., Stave, C., Sganga, A., Vincenzino, O., O'connell, B., and Stanley, R. (2005). Inequalities in dementia care across europe: key findings of the facing dementia survey. *Int. J. Clin. Pract.* 59, 8–14. doi: 10.1111/j.1368-504X.2005.00480.x

Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P., and Singh, H. (2009). Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer Dis. Assoc. Disord.* 23:306. doi: 10.1097/WAD.0b013e3181a6bebc

Cohen, T., and Pakhomov, S. (2020). "A tale of two Q15 perplexities: sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics)*, 1946–1957. doi: 10.18653/v1/2020.acl-main.176

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv [Preprint] arXiv:1712.00069.*

Duff-Canning, S., Leach, L., Stuss, D., Ngo, L., and Black, S. (2004). Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. *Neurology* 62, 556–562. doi: 10.1212/WNL.62.4.556

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49, 407–422. doi: 10.3233/JAD-150520

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Hong Kong: IEEE), 1322–1328.

Henry, J. D., Crawford, J. R., and Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia* 42, 1212–1222. doi: 10.1016/j.neuropsychologia.2004.02.001

Herd, P., Carr, D., and Roan, C. (2014). Cohort Profile: Wisconsin longitudinal study (WLS). *Int. J. Epidemiol.* 43, 34–41. doi: 10.1093/ije/dys194

Howell, K., Barnes, M., Curtis, J. R., Engelberg, R. A., Lee, R. Y., Lober, W. B., et al. (2020). "Controlling for confounding variables: accounting for dataset bias in classifying patient-provider interactions," in *Explainable AI in Healthcare and Medicine*, eds A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge (New York, NY: Springer), 271–282. doi: 10.1007/978-3-030-53352-6_25

Karlekar, S., Niu, T., and Bansal, M. (2018). Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. *arXiv [Preprint] arXiv:2006.07358.* doi: 10.18653/v1/N18-2110

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint] arXiv:1412.6980.*

Landeiro, V., and Culotta, A. (2018). Robust text classification under confounding shift. *J. Artif. Intell. Res.* 63, 391–419. doi: 10.1613/jair.1.11248

Liu, Y., and Lapata, M. (2019). "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong), 3721–3731. doi: 10.18653/v1/D19-1387

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the adress challenge. *arXiv [Preprint] arXiv:2004.06833.* doi: 10.21437/Interspeech.2020-2571

Madabushi, H. T., Kochkina, E., and Castelle, M. (2019). "Cost-sensitive bert for generalisable sentence classification with imbalanced data," in *EMNLP-IJCNLP 2019* (Hong Kong), 125. doi: 10.18653/v1/D19-5018

Marceaux, J. C., Prosje, M. A., McClure, L. A., Kana, B., Crowe, M., Kissela, B., et al. (2019). Verbal fluency in a national sample: telephone administration methods. *Int. J. Geriatr. Psychiatry* 34, 578–587. doi: 10.1002/gps.5054

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. (2013). *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, eds C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger.

Noorian, Z., Pou-Prom, C., and Rudzicz, F. (2017). On the importance of normative data in speech-based assessment. *arXiv [Preprint] arXiv:1712.00069.*

Orimaye, S. O., Wong, J. S., Golden, K. J., Wong, C. P., and Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, (Baltimore, MD) 18:34. doi: 10.1186/s12859-016-1456-0

Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 78–87. doi: 10.3115/v1/W14-3210

Orimaye, S. O., Wong, J. S.-M., and Wong, C. P. (2018). Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PLoS ONE* 13:e0205636. doi: 10.1371/journal.pone.0205636

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, eds H. M. Wallach, H. Larochelle, A. Beygelzimer, Florence d'Alché-Buc and E. B. Fox, and R. Garnett (Vancouver, BC).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Number CONF* (Barcelona: IEEE Signal Processing Society).

Rajan, K. B., Wilson, R. S., Weuve, J., Barnes, L. L., and Evans, D. A. (2015). Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia. *Neurology* 85, 898–904. doi: 10.1212/WNL.0000000000001774

Searle, T., Ibrahim, Z., and Dobson, R. (2020). Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. *arXiv [Preprint] arXiv:2006.07358.* doi: 10.21437/Interspeech.2020-2729

Stokes, L., Combes, H., and Stokes, G. (2015). The dementia diagnosis: a literature review of information, understanding, and attributions. *Psychogeriatrics* 15, 218–225. doi: 10.1111/psyg.12095

Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). "Automated screening for Alzheimer's dementia through spontaneous speech," in *INTERSPEECH*, 1–5. doi: 10.21437/Interspeech.2020-3158

Tombaugh, T. N., Kozak, J., and Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: Fas and animal naming. *Archiv. Clin. Neuropsychol.* 14, 167–177. doi: 10.1016/S0887-6177(97)00095-4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, eds I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett (Long Beach, CA).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2019). Huggingface's transformers: state-of-the-art natural language processing. *arXiv* abs/1910.03771. doi: 10.18653/v1/2020.emnlp-demos.6

Yancheva, M., and Rudzicz, F. (2016). "Vector-space topic models for detecting Alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin), Vol. 1, 2337–2346. doi: 10.18653/v1/P16-1221

Zhong, Y., and Ettinger, G. (2017). "Enlightening deep neural networks with knowledge of confounding factors," in *Proceedings of the IEEE International Conference on Computer Vision Workshops* (Venice), 1077–1086. doi: 10.1109/ICCVW.2017.131

Conflict of Interest: The 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.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 19 April 2021
doi: 10.3389/fcomp.2021.642633

Multimodal Capture of Patient Behaviour for Improved Detection of Early Dementia: Clinical Feasibility and Preliminary Results

Patrik Jonell[1†*], Birger Moëll[1†*], Krister Håkansson[2,3†], Gustav Eje Henter[1], Taras Kucherenko[4], Olga Mikheeva[4], Göran Hagman[2,3], Jasper Holleman[2,3], Miia Kivipelto[2,3], Hedvig Kjellström[4], Joakim Gustafson[1] and Jonas Beskow[1]

[1]Division of Speech, Music and Hearing, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden, [2]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden, [3]Karolinska University Hospital, Stockholm, Sweden, [4]Division of Robotics, Perception and Learning, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Kewei Chen,
Banner Alzheimer's Institute,
United States
Rubén San-Segundo,
Polytechnic University of Madrid,
Spain

Correspondence:
Patrik Jonell
pjjonell@kth.se
Birger Moell
bmoell@kth.se

†These authors share first authorship

Specialty section:
This article was submitted to Human–Media Interaction, a section of the journal Frontiers in Computer Science.

Received: *16 December 2020*
Accepted: *10 February 2021*
Published: *19 April 2021*

Citation:
Jonell P, Moëll B, Håkansson K, Henter GE, Kucherenko T, Mikheeva O, Hagman G, Holleman J, Kivipelto M, Kjellström H, Gustafson J and Beskow J (2021) Multimodal Capture of Patient Behaviour for Improved Detection of Early Dementia: Clinical Feasibility and Preliminary Results. Front. Comput. Sci. 3:642633. doi: 10.3389/fcomp.2021.642633

Non-invasive automatic screening for Alzheimer's disease has the potential to improve diagnostic accuracy while lowering healthcare costs. Previous research has shown that patterns in speech, language, gaze, and drawing can help detect early signs of cognitive decline. In this paper, we describe a highly multimodal system for unobtrusively capturing data during real clinical interviews conducted as part of cognitive assessments for Alzheimer's disease. The system uses nine different sensor devices (smartphones, a tablet, an eye tracker, a microphone array, and a wristband) to record interaction data during a specialist's first clinical interview with a patient, and is currently in use at Karolinska University Hospital in Stockholm, Sweden. Furthermore, complementary information in the form of brain imaging, psychological tests, speech therapist assessment, and clinical meta-data is also available for each patient. We detail our data-collection and analysis procedure and present preliminary findings that relate measures extracted from the multimodal recordings to clinical assessments and established biomarkers, based on data from 25 patients gathered thus far. Our findings demonstrate feasibility for our proposed methodology and indicate that the collected data can be used to improve clinical assessments of early dementia.

Keywords: Alzheimer, mild cognitive impairment, multimodal prediction, speech, gaze, pupil dilation, thermal camera, pen motion

INTRODUCTION

Alzheimer's disease and other neurocognitive disorders with a neuropathological origin develop gradually over many years before existing criteria of a clinical diagnosis are fulfilled (Blennow et al., 2006; Jack et al., 2018). The irreversible nature of these diseases and the long preclinical phase could make effective preventive non-pharmacological approaches especially appropriate, e.g., life-style changes that promote brain health and that have no negative side-effects (Kivipelto et al., 2017). Making a correct diagnosis is a challenging task, especially in early stages of these diseases (Håkansson et al., 2018); it has been estimated that more than 50% of cases of dementia are undetected (Lang et al., 2017), and that the diagnostic accuracy is only between 70 and 90%, compared to what is revealed in post-mortem neuropathology (Villemagne et al., 2018; Gauthreaux et al., 2020).

The diagnostic uncertainty in neurocognitive disorders incurs great human and monetary costs to patients and society. For the patient, a false diagnosis inflicts unnecessary trauma with devastating consequences on quality of life, in addition to medication with likely negative side-effects. For society, large cost savings are possible if only persons with a high probability of neuropathology are referred to more detailed examinations. In addition, if an underlying pathology can be correctly identified at an earlier stage, this will probably improve the efficacy of pharmacological as well as non-pharmacological counteractive measures. It is therefore of high priority to develop diagnostic tools for these diseases that are more sensitive, less invasive, more cost-effective, and easier to administer. Approaches based on machine learning have proved successful for processing complex information and assisting in medical decisions in several diseases (Hamet and Tremblay, 2017). In recent years, such methods have been developed also for neurocognitive disorders (Bruun et al., 2019; Koikkalainen et al., 2019; Lee et al., 2019a). Typically, clinical information collected through established diagnostic routines is automatically analysed, e.g., via automatic analysis of brain images. But machine learning has also been used to combine many types of clinical data to further aid in the diagnosis of neurocognitive disorders (Bruun et al., 2019; Koikkalainen et al., 2019; Lee et al., 2019a). Another potential application of machine learning for neurocognitive disorders could be the automatic capture and analysis of behavioural signals of potential clinical relevance, both for reducing the risk that such signals are missed by the clinician and for adding new and complementary information beyond what normally is collected in the medical examination. Such applications have been tested and evaluated for single digital biomarkers, such as speech or gaze, and the results have been promising in several cases, as further described in *Related Work*.

In this study we describe the first comprehensive and highly multimodal approach where signals from numerous behavioural and physiological channels are captured and analysed in parallel in real patients, as an integrated part of the regular clinical examinations at a major regional hospital. To offer a rationale for this multimodal approach, we first (in *Medical Background*) give a short medical background to neurocognitive disorders and diagnostic challenges, including neuropathological characteristics and behavioural manifestations. In *Related Work* we then describe recent developments in digital biomarkers of special relevance for this project, including speech patterns, gaze, non-verbal behaviours, and physiological signals. *Data Collection* then details our comprehensive, multimodal approach for gathering patient behaviour data during clinical interviews. This is followed by *Data Analysis*, which describes how the data can be analysed to extract digital biomarkers, and *Preliminary Findings*, which illustrates how the diagnostic relevance of the extracted biomarkers can be analysed. The implications of our preliminary findings and of our data gathering in general are discussed in *Discussion*, while *Conclusion* concludes.

MEDICAL BACKGROUND

Neurocognitive Disorders

Due to continued global increase in life expectancy, the number of persons with chronic diseases is expected to grow dramatically. As for many of these chronic diseases, age is the most important risk factor for getting a neurocognitive disorder (NCD) with a doubled risk for every 5 years of life. At the age of 90, around 50% of the population carries a dementia diagnosis, and the prevalence is around 20% higher for women than for men (Cao et al., 2020). In the case of major neurocognitive disorders (NCD), previously named dementia, no pharmacological treatment exists that can cure or halt the disease process. Approximately 50 million persons today carry some form of NCD, a number that is expected to grow to around 150 million in 2050 if no cure will be been found (Prince, 2015). Due to high-intensive need of care in later phases, these diseases put a high burden on limited care resources and societal economies. Combating these disorders has been declared a priority by the World Health Organization (World Health Organization and Alzheimer's Disease International, 2012). Neurocognitive disorders exist in various forms, where Alzheimers disease (AD) is the most common globally, accounting for approximately 60% of all cases, but limitations in vascular function to provide sufficient oxygen and nutrients to nerve cells often contribute to cognitive impairments, either alone (vascular dementia), or in parallel with e.g. AD. Cognitive disorders in older age may also derive from other neuropathological conditions such as Lewy-Body Dementia (LDB), Fronto-temporal Dementia (FTD) and Parkinson Dementia (PD), accounting in total for around 30% of all NCD cases (Cao et al., 2020). These neuropathologies are all progressive and ultimately lethal, and they typically develop during a long pre-clinical phase that, in the case of AD, may have been initiated at least a decade before diagnostic criteria are fulfilled (Jack et al., 2018). With more refined measurement techniques, including determination of various protein levels in cerebrospinal fluid and high-resolution brain imaging, it is often possible to determine which of these pathologies may lie behind also a minor NCD, previously globally referred to as "mild cognitive impairment" (MCI).

Neuropathological Characteristics and Processes

There may be several reasons for the failure to find a cure against these disorders, in spite of massive research investments across the world. The dominating disease model, on which hundreds of failed clinical trials have been based, states that AD develops through a cascade of events that are triggered by formation of beta amyloid (Aβ) protein plaques, as originally suggested by Hardy and Higgins (Hardy and Higgins, 1992). More recently, the upstream formation of neurotoxic Aβ oligomers have become more in focus than the plaques, oligomers that may later contribute to plaque formation (McGirr et al., 2020). Even if pharmacological success has been made Alzheimer's disease in terms of targeting amyloid proteins with an assumed toxicity, and even dissolving amyloid plaques, patients in these trials have not benefitted symptomatically in any of these trials (Kepp, 2017). One reason for appointing special variants of betamyloid proteins, especially the Aβ 1–42 peptide, as the culprit, is the early appearance of level increases in the brain during early phases of the neuropathological development (Long and Holtzman, 2019). But association does not prove causation, and one troubling fact for adherents of this hypothesis, besides the failures of all amyloid-based drug trials until now, is that many elderly persons have amyloid plaques, but without any clinical signs of Alzheimer's disease (Lane et al., 2018). The fact that betaamyloid accumulation does not continue to increase after the initial phase of disease development, seems to suggest that it is not directly related to the disease itself, but possibly a trigger—or even an early protective reaction against the disease (Castellani et al., 2009; Kumar et al., 2016; Li et al., 2018). As a result, doubts have been voiced against the dominating Aβ paradigm (Kepp, 2017) and other disease-related events in the brain have received increasing attention. A major alternative mechanism is related to changes in the tau protein, a building block for microtubuli, the tiny pipelines that transport substances between the soma and the synapses inside the nerve cell, but that also serve as a skeleton to maintain the structure of the cell. Degradation of the tau protein during the progression of the disease, through dysregulated phosphorylation and transformation into hyperphosphorylated proteins, makes microtubuli axonal transport progressively less efficient, leads to synapse loss, to formation of neurofibrillary tangles (NFT) and ultimately cell death. Some findings indicate that these changes start in very early stages of disease development, even before changes in Ab (Insel et al., 2020). In contrast to Aβ changes, degradation of tau progresses further in parallel with the disease (Long and Holtzman, 2019) and may therefore be a better indicator of disease stage, compared to measures of Aβ (Lane et al., 2018). Changes in Ab and tau proteins are often seen as related, and, according to advocates of the betaamyloid cascade hypothesis, changes in extracellular Aβ precede and trigger tau hyperphosphorylation inside the neuron (Phillips et al., 2020); a detailed diagnostic evaluation typically involves measurement of both these proteins in cerebrospinal fluid, especially levels of the Aβ 1–42 molecule and levels of total tau and phosphorylated tau (p-tau). The coexistence of extracellular accumulation of beta-amyloid and the development of neurofibrillary tangles (NFT) are still considered as the main pathological markers of AD, but no drug trials based on either of these targets have so far been successful (Long and Holtzman, 2019). Other suggested mechanisms include cholinergic deficits, evidenced by the relative efficacy of cholinesterase inhibitors to hamper cognitive decline in AD (Sharma, 2019), and inflammation, indicated by microglia and astrocyte activation in AD.

Behavioural Manifestations

Whatever the mechanisms behind, established effects on cognition (Henneges et al., 2016) and on behaviour seem logical from what we know about the underlying pathology and its progression. Usually these pathological changes in AD start in the medial temporal part of the brain, from where it propagates to neighbouring areas, and to areas with projections from already affected areas. As this part of the brain, including the hippocampus and entorhinal cortex, has a central role for especially working memory and episodic memory, these functions are typically affected in early phases, albeit subtly at first. The olfactory bulbs are close neighbours, and impaired olfaction is also a typical early sign (Phillips et al., 2020).

Both the ability to understand language and to speak have important centres in the parieto-temporal and the temporal lobe, and are also typically affected relatively early, and could lead to slower and less articulated speech, difficulty in finding words, and difficulties to understand language. These functions are normally controlled from the left hemisphere, while the right parieto-temporal hemisphere is relatively more important for spatial functions and orientation. Difficulty in drawing figures and navigation are common behavioural manifestations that most probably are related to impaired function in this part of the brain, in combination with impairments in especially the enthorhinal cortex. Decreasing efficiency of neural functional (e.g. in axonal transport, transmitter substance deficits, and an impoverished synaptic network and neural interconnectivity) will also have a number of more general effects that in a progressive manners will affect associative ability, reaction time, balance and motor coordination. When the neuropathology spreads further, impulse control, attention, and the ability to focus are affected, mainly regulated by the fronto-temporal lobes (Migliaccio et al., 2020).

Long-term memory, especially procedural memory, are spared until late in the pathological development, indicating less importance of parieto-temporal regions for these functions. The different effects on short term vs. long term memory is often illustrated by the ability to detail events that happened decades ago, while the person may have no recollection of what happened earlier the same day or week. For example, patient with clinical AD may not remember that he or she can play the piano, but positioned in front of one, could still start to play it. Recently it has been suggested that the typical AD phenotype is not the only one, and what we call Alzheimer's disease should be considered as a family of related diseases, but with important differences in neuropathology, e.g., in terms of primarily affected areas and thereby also in cognitive and behavioural manifestations and the sequence of their appearance (Ferreira et al., 2017). The progressive nature of AD and other neuropathological diseases means that eventually the whole brain will be severely affected and thereby all cognitive and behavioural functions. As a result, dementia care in late stages is resource demanding and, in combination with increasing longevity and the high prevalence in old age, presents a large and growing economic burden for societies worldwide (Wimo et al., 2017).

Assumptions and Rationale for This Project

It seems plausible that odds would improve with earlier intervention for any strategy against any disease, including both pharmacological and non-pharmacological strategies, as long as it is based on an adequate assumption of the underlying disease mechanism. There are however special challenges with AD and other neuropathologies leading to NCD, due to a very long progressive disease development with subtle symptoms in the earliest stages. The limited therapeutic success against AD and other neuropathological diseases indicates that the underlying mechanisms are not yet fully understood, which could justify a broad, open and non-biased approach. A fundamental starting point for such a non-biased and exploratory approach is the assumption of a link between brain and behaviour; we know for sure that these diseases are diseases of the brain, and this means that aspects of behaviour related to affected brain areas also should be affected, albeit subtly in early stages. To exemplify, episodic memory is typically affected in AD, most probably due to early damages to hippocampal and entorhinal regions. It could be assumed that this cognitive domain is also subtly affected in very early stages, but may not easily be captured by test scores in existing cognitive tests. But even if actual test scores should appear non-indicative of an existing neuropathology, the subtly affected person may still feel more anxious and need to make more of an effort to perform at this level, which should reflect in various ways in the behaviour of the person, not easily detected by the naked eye. The same principle should apply to any other cognitive domain that has been subtly affected, whether it be reading ability, executive functioning, word finding, or processing speed, depending on the type of neuropathology and which brain areas are affected by it. Another example is autonomic function that typically has a lower range of variability, being "flatter", if a person is carrying a neuropathological disease (Algotsson et al., 1995). Autonomic function should reflect in degrees of heart rate variability, variability in emotional expressions, skin temperature fluctuations, speech volume variation, and in pupil size variations. Could any or several of these indicators be identified in early stages and will they differ between different types of NCD?

In this project we use a broad approach to automatically and continuously capture a large number of potential digital biomarkers with high precision, by using different sensors. We then subject the collected data to machine learning to identify signals and patterns of signals that could indicate an underlying neuropathology. In the following we will in greater detail describe the rationale behind each type of potential digital biomarker that we capture.

RELATED WORK

This section explores how related sensor data, and digital biomarkers extracted from such data, across different modalities have previously been considered for clinical assessment of Alzheimer's disease.

Digital Biomarkers

The term digital biomarkers is used here to specify metrics extracted from sensor data and differentiate them from biological biomarkers extracted from biological measurements. A digital biomarker reflects the underlying state of the biological system (the human brain) and a good candidate for a digital biomarker is one that shows promise in identifying both diagnostic criteria of AD and correlates with established biomarkers used in AD examination. This section outlines what digital biomarkers have been used in previous research. All digital biomarkers used throughout this article are written in *italics*.

Speech and Language

Alzheimer's disease leads to a decline in cognitive and functional abilities, such as memory loss and language impairments. There have been numerous review studies on linguistic biomarkers that have been used for detecting the progression of AD (Mueller et al., 2018; Slegers et al., 2018; Voleti et al., 2019; de la Fuente Garcia et al., 2020; Calzà et al., 2020). These include both acoustic features (prosodic, spectral, vocal and fluency), and textual features (lexical, syntactic, semantic, and pragmatic). Vocal features such as *speaking rate*, *fluency* and *voice quality* could be useful as biomarkers for early detection of AD, since they stem from atrophy in the medial temporal lobe (König et al., 2015). In a longitudinal study Ahmed et al. (2013) found that lexical, syntactic and semantic complexity changed significantly as the the disease progressed, but not voice quality or fluency. Speaking rate have been found to be the earliest measurable linguistic feature for AD detection (Szatloczki et al., 2015). MCI patients have been found to have a more breathy (H1-A3) and weaker voice (CPP) than NC (Themistocleous et al., 2020). *Number of silent pauses* (especially those longer than 2 s) have proven to be useful for AD detection (Yuan et al., 2020), as has the *average length of silent pauses* (Roark et al., 2011; Tóth et al., 2018). The increase in pause frequencies has been attributed to struggles with lexical retrieval, but might also reflect other cognitive impairments as pauses increases with cognitive load (Pistono et al., 2016). In a study on language use in unstructured interviews, AD subjects were found to use fewer Nouns, while more Adjectives, Verbs and Pronouns than healthy older participants. They also used a smaller *vocabulary size* (Bucks et al., 2000). The lexico-semantic variables appear to be the most useful for the diagnosis of later stages of AD (Boschi et al., 2017). These results suggest that the occurrence of dementia is associated to reduced syntactic complexity, difficulty in connecting one event to the next, in maintaining the theme, and in understanding the story. Furthermore, grammatical errors have mainly been observed in severe AD groups (Jarrold et al., 2014). Some semantic features seem to be relevant for MCI though. Asgari et al. (2017) tagged transcription of patient doctor interviews using the Linguistic Inquiry and Word Count (LIWC). Using this, they divided the words into five broad categories: Linguistic processes; Personal concerns, Psychological processes; Relativity and Spoken categories. The category that was most significant for MCI was the relativity category that included words dealing with time and space. Haider et al. (2019) demonstrated the usefulness of purely acoustic features, e.g. eGeMAPS (Eyben et al., 2015), openSmile (Eyben et al., 2010), and ComParE (Eyben et al., 2013), that has proven useful for other paralinguistic detection tasks.

Facial Gestures

The effects of AD on facial gesture and expressiveness can be significant, but it is a complex relationship. Overall facial biomarkers are most related to the later stages of AD with the MCI group having different facial expression in relation to the AD group. On the one hand, apathy is one of the most common behavioural symptoms of AD and is linked to deficits in goal-directed behaviour, decreased goal-related thought content and emotional indifference with flat affect (Cai et al., 2020), which in turn leads to overall reduced facial expressivity (Seidl et al., 2012). Asplund et al. (1991) found that patients in the later stages of AD struggled to show *facial emotional reactions* when experiencing emotional stimuli. Burton and Kaszniak (2006) found reduced correlation between emotional state (valence) and zygomatic activity (smiling) for patients with AD. The AD patients experience the emotion (happiness) but are less likely to do the linked zygomatic activity (smile). On the other hand, dementia is also generally linked to reduced control over facial expression, in many cases leading to *increased* facial expressiveness. Smith (1995) found that people with mild dementia exhibited reduced control of negative expression during a picture stimuli experiment. The relationship between stimuli and facial muscle expression of emotion is complicated since deficit in emotional facial expression can be caused by several factors. Seidl et al. (2012) concluded that cognitive deficits are associated with increased rate of total facial expression after controlling for apathy. In addition, Matsushita et al. (2018) found that AD patients had an increased tendency to use smile as a "save appearance response" when they fail to provide the correct answer to questions.

Motor Signs (Hand and pen Motion)

Even though cognitive impairments are the most common signs of dementia, motor functions are also affected by the disease. Motor signs like speech/facial expression, rigidity, posture, gait and bradykinesia have been found to increase in frequency and severity over time in AD patients (Scarmeas et al., 2004). Chung et al. (2012) has developed an inertial-sensor-based wearable and a stride detection algorithm for analysis of Alzheimer patients' gait behaviour. In a user study they were able to show difference in gait profiles between the AD patients and the healthy controls. The finger tapping test is used as a neuropsychological assessment of fine motor skills (Reitan and Wolfson, 1985). It has been found useful for AD assessment, where AD patients produced a finger tapping pattern that was lower in frequency with slower, more variable inter-tap interval than the health control group (Roalf et al., 2018). Previous studies show that MCI and AD patient have a lower *drawing speed* when performing handwriting tasks with lower *pen pressure* with the differences corresponding to the groups with more deteriorated groups showing larger differences. Only using these kinematic measures, a classification accuracy of 69–72% was achieved. (Werner et al., 2006). Gatouillat et al. (2017) propose some novel measurements/features: pen-tip normal force, total grip force, and an objective writing quality assessment. They do not correlate with cognitive aspects per se, but measure trade-offs between timing and accuracy in the writing and such things. Garre-Olmo et al. (2017) used a digital pen in a number of tasks (Clock test, copying two and-three dimensions drawings, copying one sentence, writing dictated sentence). Apart from speed and pressure, they found that the time the pen was in the air was a discriminant feature between AD, MCI and NC.

Gaze and Pupil Dilation

There has been research on understanding cognitive deterioration and dementia from *eye movements* (Zhang et al., 2016). For different tasks, the *eye movements* of people with AD differs from control subjects (Beltrán et al., 2018). Gaze patterns of patients with AD show greater variance in all directions. This is linked to cognitive decline and deficits in attention which leads to more frequent eye and facial movement (Nam et al., 2020). AD patients have also been found to have problems following a moving target (Molitor et al., 2015). These variations in gaze in AD patients are likely due to damage to frontal and parietal lobe regions related to attention (Garbutt et al., 2008). When comparing facial muscles and eye movement, less variability is seen for AD patients compared to healthy controls (Nam et al., 2020). *Pupil dilation* is a robust predictor of cognitive load, the working memory demands of performing a certain task (Gavas et al., 2017). *Pupillary response*, mainly in terms of changes in reaction to light, has been proposed as a biomarker of early stages for Alzheimer's disease (Granholm et al., 2017), However, a longitudinal study with AD biomarkers is needed to confirm whether pupillary responses can provide a predictive biomarker of risk specific to AD-related declines.

Autonomic Nervous System

Heart rate variability (HRV) has been used extensively to predict dementia (Allan et al., 2005; Zulli et al., 2005; Negami et al., 2013) as was recently reviewed in da Silva et al. (2018). There is no consensus in the field, as some studies found that HRV time and domain parameters were lower in patients with AD than in patients with MCI and controls (Zulli et al., 2005; de Vilhena Toledo and Junqueira, 2010), while others found no difference (Wang et al., 1994; Allan et al., 2005). In general, there is no strong evidence to use of the HRV alone as biomarkers to diagnose dementia (da Silva et al., 2018). The sympathetic nervous system can also be probed using a Galvanic Skin Response sensor, such as the Empatica wrtistband, has been found to be useful in determining stress during activities (Schlink et al., 2017). *Sympathetic skin response* (SSR) and HRV together were used to detect an abnormality of autonomic function in patients with AD (Negami et al., 2013).

Thermal Emission

Experiments on using Thermal imaging for inferring stress indicate a relationship between an increase of workload and thermal emissions (Anzengruber and Riener, 2012). Zhou et al. (2019) used a wearable thermal sensor and found that it can be possible to use such a system for estimating mental workload. Ruminski and Kwasniewska (2017) presents a review of thermal imaging in mobile conditions together with a proposed prototype. Furthermore, sleep-disordered *breathing* is associated with a higher risk of AD onset after matching and adjusting for other risk factors (Lee et al., 2019b). Recent pilot study, Tiele et al. (2020) confirms the potential utility of analysing breath volatile organic compounds to distinguish between MCI, AD and controls. Respiration rate has successfully been extracted from thermal imaging by automatically analysing the thermal fluctuations in the nostril area (Lewis et al., 2011). Cho (2018) used a mobile thermal imaging device in order to infer "stress" levels by extracting respiration rate.

Automatic Capture and Analysis of Cognitive Assessment Tests

Recently, there have been large efforts in automating the screening of Alzheimer's disease. Tóth et al. (2015) report a completely automated speech-based screening pipeline that yielded significant discrimination results. König et al. (2018) has developed an iPad application that can perform a semantic verbal fluency test and automatically perform a fine-grained analysis of the spoken input. ICAT is an internet-based cognitive assessment tool that uses speech recognition for a delayed list learning task and drag and drop GUI input for a number sorting task (Hafiz et al., 2019). In the Talk2Me project anonymous people can contribute with both speech and text via a web interface (Komeili et al., 2019). The speech tasks include describing a picture and retelling a story that is displayed on the screen for a short while. The text-input tasks include image naming, word naming and providing word definitions. The authors have also developed a linguistic analysis package called COVFEFE that they have made available as open source. Intelligent Virtual Agents have also been used to collect spoken interactions, for example to automate parts of the initial interview at a memory clinic Mirheidari et al. (2017). In a series of studies the team has used a mix of automatically generated acoustic and lexical features with manually acquired conversational analysis inspired features to predict AD (Mirheidari et al., 2019; Walker et al., 2020). Today's smart phones and wearables have a large number of sensors that could be used in data collection for dementia detection. This includes camera, microphone, accelerometer/gyryscope, touch, geoposition, ECG and IR cameras (Kourtis et al., 2019). Using wearable consumer products have been used for continuous monitoring of symptoms related to cognitive impairment (Chen et al., 2019). As an example, UbiCAT is a ubiquitous cognitive assessment tool for smart watches, that includes three cognitive tests: the Arrow two-choice reaction-time test, the N-back letter test, and the Stroop color-word test (Hafiz and Bardram, 2020).

In the current study we present a multimodal capture and analysis framework that makes use of non-obtrusive and affordable sensors in capturing the human behaviour during memory tests. It has been integrated into the fast-track cognitive assessment procedure that is used at the memory clinic of a major regional hospital in Sweden.

DATA COLLECTION

We now describe the setup and procedures we used for gathering our multimodal behavioural and phsyiological data. All recordings were performed during clinical examinations at the Memory Clinic at Karolinska Hospital in Stockholm, Sweden. The examinations are part of an established fast-track analysis where a multi-disciplinary team assess the patient within one week. The complete examination includes brain scanning (MRI), neuropsychological assessment, speech and language assessment, assessment of motor skills, physical examination, and a 1-h clinical interview. Our recordings took place during the clinical-interview portion of the examinations, the procedure of which was minimally modified and standardised to accommodate the recordings, as described in *Procedure*.

During most of the clinical assessments at the clinic the patient and the clinician are sitting on opposite sides of a table. In some cases, including some of our recordings, a partner or relative of the patient may be present and sitting beside the patient. For our study, these assessments took place in a particular room at the clinic, where the room and the table had been instrumented for multimodal data capture. **Figure 1** shows the custom-built, instrumented "recording table" used. The entire setup encompasses *sensors* for recording, *interfaces* for controlling, monitoring, and performing data gathering, along with miscellaneous *other equipment*, e.g., for storing the data, and a *recording software infrastructure* that coordinates the different devices and ties everything together. The remainder of this section describes the various components in more detail, along with the procedures for conducting the clinical sessions and exporting the data. For an overview of what modalities each sensor captured, please see **Table 1**. **Figure 1** shows the data collection setup from the clinical environment. The clinical assessments at the hospital conclude with a physical examination in a different part of the room, but this part of the assessment procedure was not recorded, since the potential added benefits of such data was not considered commensurate to the privacy intrusion it would entail.

Design Considerations

A key consideration when designing the data-collection methodology was to create a setup with a minimal impact on the clinical assessment, in order to maintain the ecological validity of the collected corpus. For example, eye movements and pupil dilation can be collected either using a display-mounted eye tracker or by having the user wear eye-tracking glasses. Although the glasses are much more effective, they are cumbersome to wear, distractive, and also increase the sense of being monitored. We therefore opted for a display-mounted eye tracker instead. The case of audio recording is similar: a head-mounted microphone provides better quality than microphones fixed to the table generally do, but again, requires equipping the patient with hardware. Considering these facts, we settled on using a setup with mobile phones (Apple iPhones) mounted to the table, which are less associated with looking like cameras than other types of "normal" cameras, for capturing video and facial data. We also use an array microphone integrated into the table

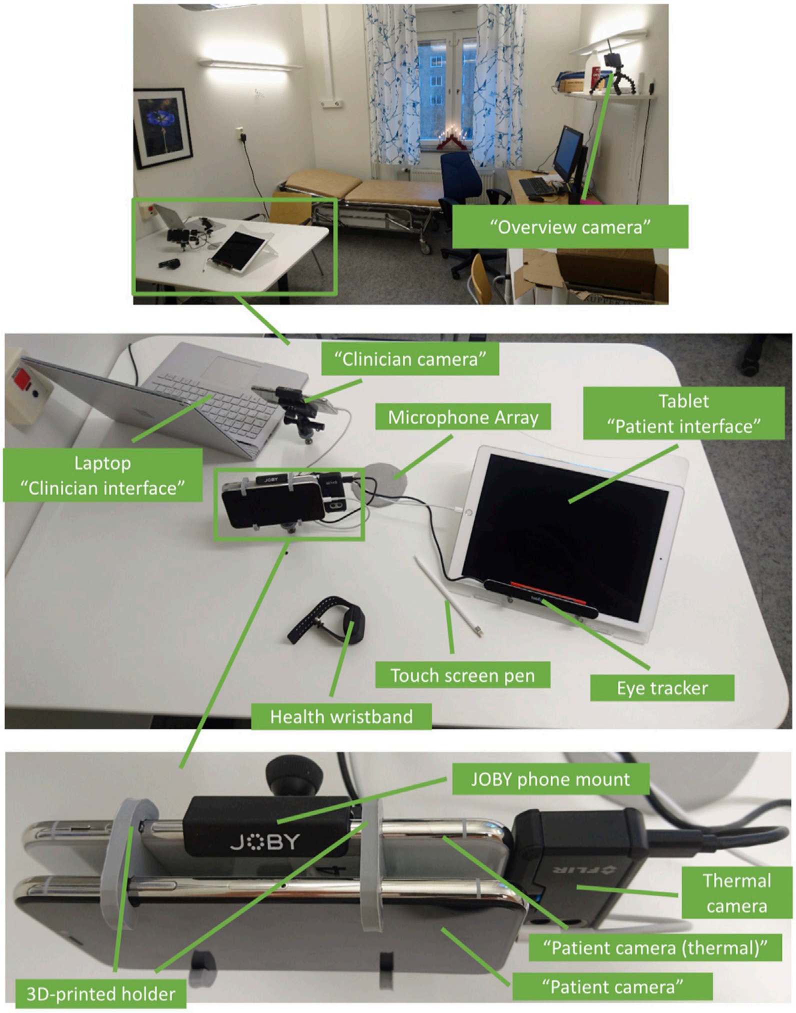


FIGURE 1 | The data collection setup. At the top an overview of the room is given, showing both the instrumented recording table and the position of the "overview camera". In the middle the various devices on the instrumented recording table are shown and at the bottom a close-up of the patient facing cameras.

which is able to capture speech from both the clinician and the patient. For eye-tracking we opted to use a Tobii Nano which is able to capture eye movement and pupil dilation at a distance, attached to the bottom of the tablet. The only device which the patient is carrying is a health wristband, which was considered to not be as invasive, since it is not uncommon to wear a watch on the wrist.

Sensors

Below we introduce the various sensors and equipment used for the data collection procedure (**Table 2**).

Cameras

Similar to Malisz et al. (2019) a pair of Apple iPhones X (from here on referred to as "Patient camera" and "Clinician camera") were used in order to record both the patient and the clinician. An additional, third iPhone X was used for capturing thermal data ["Patient camera (thermal)"] from the patient, and a fourth capturing the whole interaction from a distance ("Overview camera"). Please see **Figure 2** to see how the iPhones were connected with the system, and **Figure 1** to see how the cameras were placed and mounted. For the three iPhones capturing close-ups of the patient and clinician ("Patient iPhone", "Patient camera (thermal)", and "Clinician camera"), a mount from JOBY was modified and attached to the table. Furthermore a holder was 3D-printed in order to attach the "Patient camera" with the "Patient camera (thermal)" (see **Figure 1**). As can be seen in **Figure 1** the "Patient camera (thermal)" had a FLIR One thermal camera attached to it, together with a charging cable. These iPhones used a software developed specifically for these data recordings, and synchronised their time with the FARMI server. When starting the application all the recording options were presented, and which data streams that should be captured could be selected. Those were; RGB video, facial gestures (parametrised facial expressions and head movement), depth data, 3D-mesh data, thermal video, RGB reference video for the thermal video, and thermal data. As can be seen in **Figure 3**, the various data streams can be

TABLE 1 | A summary table of what modalities each sensor captures.

Sensor	Modality	Captures
Eye tracker (tobii nano)	Gaze	Patient
	Pupil dilation	Patient
Health wristband (empatica E4)	Heart rate	Patient
	Galvanic skin response	Patient
	Accelorometer	Patient
Cameras (4 apple iPhone +1 FLIR one)	Video	Patient, clinician, and overview
	Facial gestures	Patient, clinician
	Thermal emission	Patient
	Voice	Patient, clinician
Microphone Array (ReSpeaker mic array v2.0)	Voice	Patient, clinician
	Language	Patient, clinician
Tablet (Apple iPad)	Pen movement	Patient
	Pen pressure	Patient

turned on or off. The iPhones were configured to send out an image every 3 s which the status page could display, in order for the technician to act in case there were issues with the video.

Health Wristband

Originally an Apple Watch was used in order to capture heart rate and accelerometer data for the patients. The apple watch was later replaced with an Empatica E4 wristband that captures heart rate, accelerometer data, and electrodermal activity.

Microphone Array

A microphone array (ReSpeaker Mic Array v2.0) was installed into the table in an approximately 10 cm round hole in the center of the table. The microphone array was covered with a mesh cloth (see **Figure 1**). The microphone array was connected using a USB cable to the central computer. The default LED lights indicating the direction of speech were disabled, as they were deemed distracting.

Eye Tracker

A Tobii Nano was used in order to capture eye movement and pupil dilation of the patient while interacting with the Tablet. **Figure 1** shows how the eye tracker was placed. A custom mount for the tablet was 3D-printed in order to place the eye tracker at an appropriate height and angle with respect to how the patient sits. A manual calibration procedure was required before each session, where the patient was asked to focus their gaze at circles displayed on the tablet. The calibration was initiated from the status page and performed together with a technician. The eye tracker was connected to the central computer. The eye tracker collected data throughout the whole assessment but was meant primarily for when the patient interacted with the tabled.

Tablet

A tablet was used (Apple iPad) together with a touch enabled pen (Apple Pencil) which hosted the clinician interface (described in *Clinician Interface*). The tablet was placed in a stand with some inclination (see **Figure 1**) such that it would be easily operated for the patient without the need of moving the tablet.

Interfaces

There were three user interfaces, one for the patient, one for the clinician, and a monitoring tool for monitoring the session. All of the user interfaces were web applications which were hosted on the central computer. Each of them are described below.

Patient Interface

A tablet interface was developed to replace certain parts of the MOCA test. The tablet interface was a web interface controlled by the clinicians interface (described below) and was black when nothing was displayed in order to not to be distracting. The tablet was used for six tasks:

- Cookie theft test, where the participant was presented an image and asked to describe what they see.
- Cube drawing, where the participant is asked to draw a copy of a three-dimensional cube which is presented to them.
- Three images, where the participant is presented with three images, and asked to describe them
- Trail making test (TMT), where the participant is presented with a number of letters and numbers, and asked to trace a line between them in ascending order alternating between letter and number each time (1, A, 2, B . . .).
- Clock drawing, where the participant is asked to draw a clock, with the time set to ten after eleven.

For the tasks were the patient had to input something (Cube drawing, TMT, and Clock test) the interactions were performed using an Apple Pencil, and all movements together with the pressure applied when drawing was recorded.

Clinician Interface

The clinician interface (see **Figure 4**) was a web application displayed through a touch-enabled laptop (Microsoft Surface). The clinician was able to choose what was displayed on the tablet interface for the patient, or just to make the patient screen go blank. It was also possible for the clinician to end the recording from this interface. The clinician also received the results from the drawing tasks through this interface, as the tablet was positioned toward the patient. These drawings could then be printed and added to the patients medical journal.

TABLE 2 | A summary table of what physiological and behavioural measures can be extracted from each modality, an indication of which ones are used in the correlation analysis and an indication if the measure is task independent.

Modality	Measure	Part of preliminary analysis	Task independent
Facial gestures	Mean face velocity	✓	✓
	Mean smile	✓	✓
	Mean brow	✓	✓
	Mean jaw	✓	✓
	Head motion	✓	✓
	Facial gaze measurements	✓	✓
	Facial patterns	✗	✓
	Emotion expression	✗	✓
Gaze	Number of fixations	✓	✗
	Mean fixation duration	✓	✗
	Number of reading fixations	✓	✗
	Number of reading backtrack	✓	✗
	Percent backtrack	✓	✗
Hand motion	Gait	✗	✓
	Hand movement	✗	✗
Heart rate	Heart rate variability	✓	✓
	Heart rate change over time	✗	✓
Language	Average word length	✓	✓
	Unique words	✓	✓
	Part-of-speech-tagging	✓	✓
	Word complexity	✗	✓
	TFIDF-vectors	✗	✓
Pen motion & pressure	Drawing speed	✓	✗
	Pen pressure	✓	✗
Pupil dilation	Pupil change	✓	✗
	Pupil diameter	✓	✗
Galvanic skin response	Electro-dermal activity	✗	✓
Thermal emission	Head temperature change	✓	✓
	Breathing	✗	✓
Video	Skin color changes over time	✗	✓
	Posture	✗	✓
	Body movement	✗	✓
Voice	h1h2 (voice quality)	✓	✓
	h1h3 (voice quality)	✓	✓
	h1a1 (voice quality)	✓	✓
	h1a2 (voice quality)	✓	✓
	h1a3 (voice quality)	✓	✓
	Average pause length	✓	✓
	Mean long pause length	✓	✓
	Pause count	✓	✓

Monitoring Tool

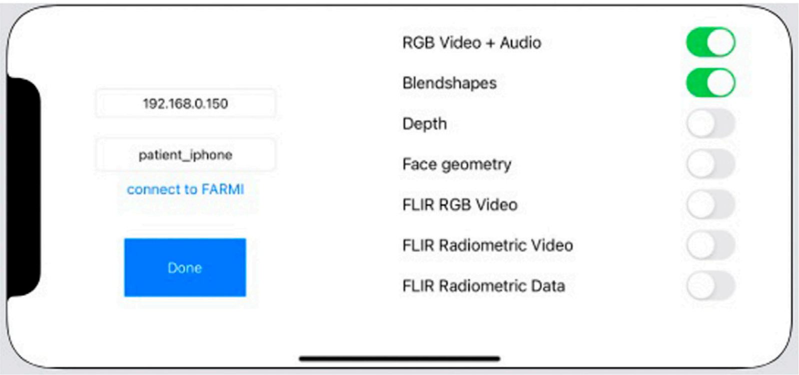
A monitoring tool in the form of a web application was created in order to be able to monitor the recordings (see **Figure 5**). Each sensor except the wristband sent a "heartbeat" signal with an interval of 5 s to the recording server (described below). This heartbeat was used in order to determine whether a device was connected to the recording setup or not, and displayed as a red or green indicator on the status page. Furthermore a still image captured by the iPhones every 3 s was also shown on the status page in order to see that data is being collected accordingly. Statistics about memory and processing usages, and battery information for the FLIR One camera was also presented. The status page was used to start and stop the recordings, and also initiated the eye-tracking calibration on the patient interface.

Recording Software Infrastructure

Since the aim was to have a recording setup with a large number of sensors, computers, mobile devices and wearables working together, it was of central importance to have a communication framework that would allow for a finely controlled synchronisation of all data streams and remote access to start and stop recordings across the various devices involved. To accomplish this, we used a modified version of the open-source FARMI framework[1] for recording multimodal interactions (Jonell et al., 2018).

The different devices used for the recordings provide data streams of different frame rates, and each device has its own internal system time that is likely to differ between devices. FARMI was designed to synchronise such streams in a robust

[1] https://github.com/kth-social-robotics/multisensoryprocessing

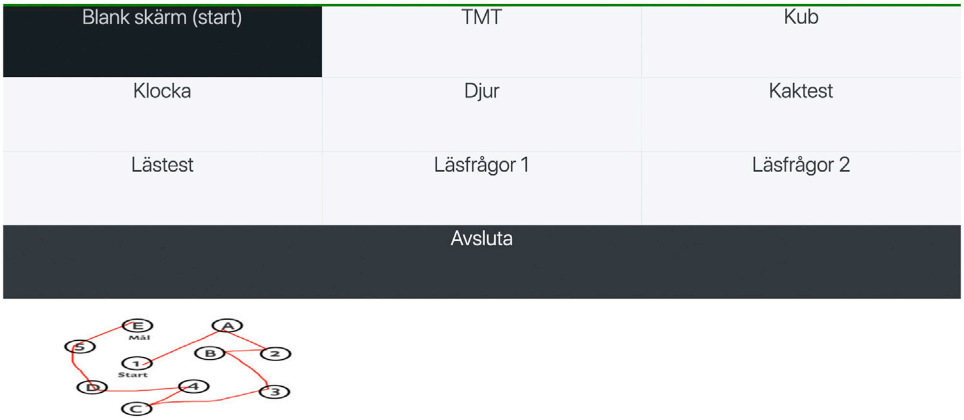


192.168.0.150

patient_iphone

- RGB Video + Audio ☑
- Blendshapes ☑
- Depth ☐
- Face geometry ☐
- FLIR RGB Video ☐
- FLIR Radiometric Video ☐
- FLIR Radiometric Data ☐

FIGURE 2 | Diagram showing each sensor component and how they are linked together with the data capturing framework.



Blank skärm (start)	TMT	Kub
Klocka	Djur	Kaktest
Lästest	Läsfrågor 1	Läsfrågor 2
Avsluta		

FIGURE 3 | The interface used to set up the iPhones before a data capturing session. Here one can set the IP address and an identifying name of the phone. Furthermore one can select which data streams to capture.

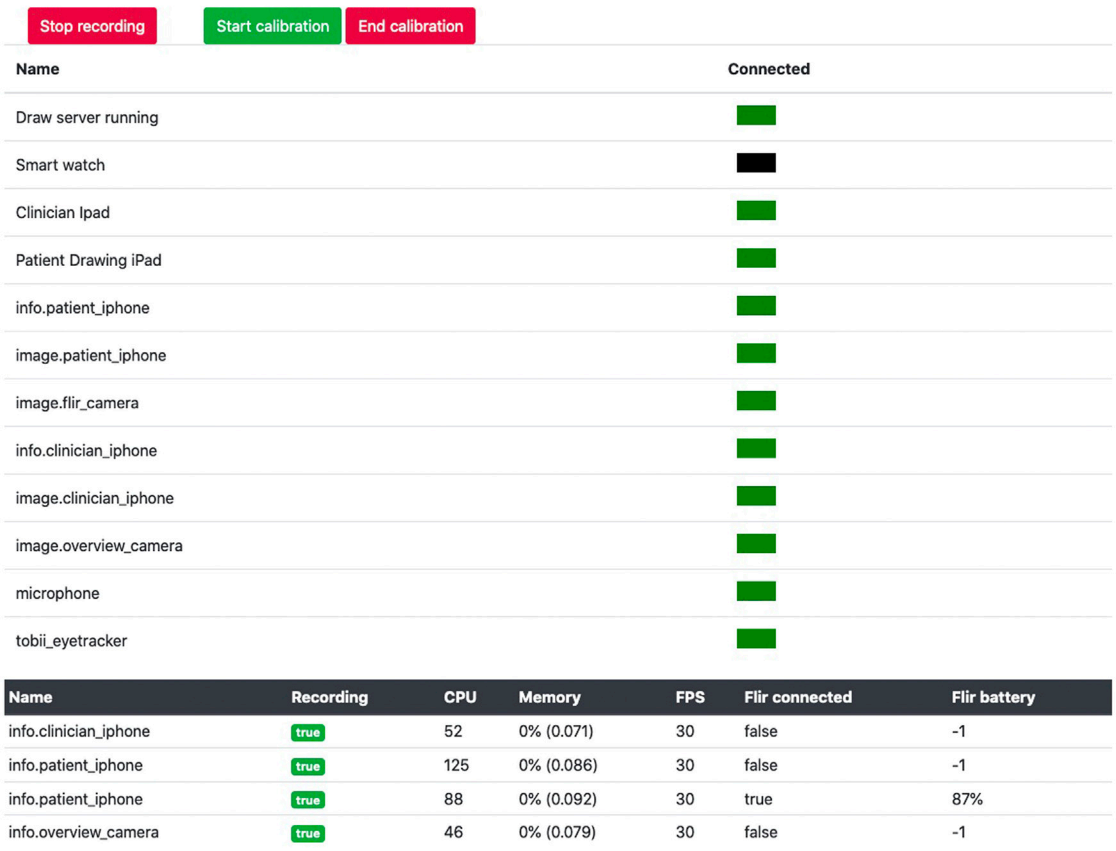
manner. It acts like a publish–subscribe framework, meaning that components in the system can either publish data at a certain topic or subscribe to receive data from a certain topic, and ensures that each device always has a known time offset relative to a central server, and that each data packet which is stored or sent out is timestamped with a timestamp synchronised with that central server. The overall software architecture is illustrated in **Figure 2**. It is a decentralised system where each component works independently of the other. Three publish-subscribe topics were used, one named "Start-Stop", which was used for sending out a signal to all devices to start recording, one named "Status image/info", which the cameras used to send a an image every 3 s to the monitoring tool along with various usage statistics, and lastly a heartbeat topic which was used by all devices to signal to FARMI that the devices were still operational.

Besides being a framework, FARMI also provides a server. Specifically, each sensor or interface would start a ZeroMQ[2] server, and send their IP addresses together with a topic name to the central FARMI server. This server would then be used as a directory service by other parts of the network for knowing which IP address a certain type of data was being published at. When a new sensor connected to the framework, this information was sent to all other connected devices so that they could connect to the new device if needed and subscribe to its data stream(s). To verify that they were still operating correctly, all sensors also published a so-called "heartbeat" signal at 5 s intervals that the FARMI server subscribed to. This was used to remove entries in the directory that had not properly sent an explicit shutdown signal to the server.

The different interfaces used to control, monitor, and carry out recording also leveraged FARMI. Specifically, each of the the patient interface, clinician interface, and monitoring tool was a web interfaces hosted on the central computer named "Web server" in **Figure 2**. The clinician interface could control what was shown on the patient interface, through communication via websockets[3]. Both the clinician interface and the monitoring

[2] https://zeromq.org/

[3] https://developer.mozilla.org/en-US/docs/Web/API/WebSockets_API



Name	Connected
Draw server running	🟩
Smart watch	⬛
Clinician Ipad	🟩
Patient Drawing iPad	🟩
info.patient_iphone	🟩
image.patient_iphone	🟩
image.flir_camera	🟩
info.clinician_iphone	🟩
image.clinician_iphone	🟩
image.overview_camera	🟩
microphone	🟩
tobii_eyetracker	🟩

Name	Recording	CPU	Memory	FPS	Flir connected	Flir battery
info.clinician_iphone	true	52	0% (0.071)	30	false	-1
info.patient_iphone	true	125	0% (0.086)	30	false	-1
info.patient_iphone	true	88	0% (0.092)	30	true	87%
info.overview_camera	true	46	0% (0.079)	30	false	-1

FIGURE 4 | The clinician interface (in Swedish). The patient has just performed the TMT test, and drawn the connecting lines. The clinician has then made the screen blank. The interface is designed to be operated through a touch screen.

tool could send out a start or stop signal via the FARMI Start-Stop topic. Furthermore, the monitoring tool could instantiate calibration of the eye tracker, and would at the same send a signal via websockets to the patient interface to show the eye-tracker calibration screen.

Most of the software connecting the sensors with the central computer was written using Python and the FARMI framework, however the code for the cameras, which were Apple iPhones, was written in Swift, utilising the FLIR framework[4] for thermal images, the ARKit framework[5] for capturing facial gestures and video, and the FARMI framework for communication with other devices. Sound was also recorded. This data was then stored locally on the phone, but timestamped using synchronised timestamps from the FARMI framework. Images and phone health statistics were published using FARMI every third second in order to be displayed on the monitoring interface. All sensors subscribed to the Start-Stop topic in order to receive a signal when to start and stop recordings. The gaze recorder used the Tobii SDK[6] to communicate with the Tobii Nano device, while the audio recorder used a Python library from ReSpeaker[7] to communicate with the microphone array.

Other Equipment

A printer was used for the clinicians to print out the results from the MOCA test for purposes of medical record keeping. The printer was connected via WiFi to the router, and could be accessed from the clinician's computer. A router (Asus RT-AC66U) was used to connect all the devices. For data security, this router was not connected to the Internet, meaning that the entire data-collection setup was isolated from the Internet. A Bluetooth-connected button was initially used for capturing points of interests deemed by the clinician during the recording sessions. This turned out to be difficult to maintain, and is thus not part of the final dataset.

Procedure

In this section we describe the procedure of the data capture from selection of patients to recordings during clinical assessments, data export and collection of biomarkers.

[4] https://developer.flir.com/mobile/flironesdk/

[5] https://developer.apple.com

[6] http://developer.tobiipro.com/python.html

[7] https://github.com/respeaker/respeaker_python_library

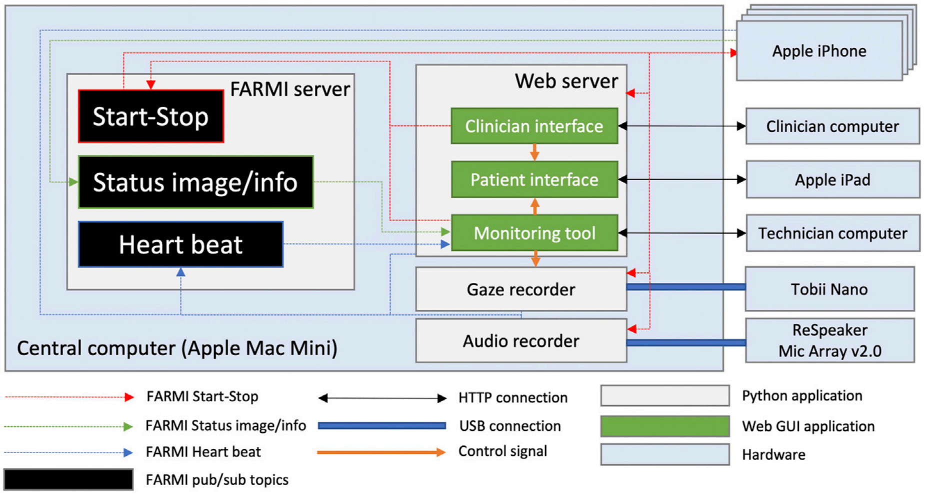


FIGURE 5 | Interface of the monitoring tool used by the technician. The images from the cameras have been cropped out for privacy reasons.

Selection and Recruitment of Participants

The participants in this study are recruited among patients at the Memory Clinic at Karolinska University Hospital in Solna, Sweden. The clinic specialises in relatively young patients with cognitive complaints, and many the patients are referred from other clinics to receive a thorough and advanced evaluation. The prevalence of dementia is below 1% for persons between 60 and 65 in all parts of the world (Ferri et al., 2005) and a dementia diagnosis below the age of 55 is very rare. Persons below 55 years of age were therefore excluded for reasons of clinical relevance and generalisability. To avoid expectation effects on patient behaviour in the interview situation, patients with an obvious or very probable neurocognitive disorder, as revealed by referral medical documentation, were also excluded. To reduce variability from interviewer behaviour, almost all interviews are carried out by one of two physicians who were trained to perform the examination to fit the requirements of the study (including use of tablets instead of paper and pen in some tasks, positioning of chairs for optimal video capture, and administration of additional tasks, as described above).

At this point, we have recorded 25 patients before the outbreak of the COVID-19 pandemic suspended the data gathering, with our aim being to recruit and record 100 patients in total. Based on previous data from the clinic, we expect that approximately 50% of these will be diagnosed with a neurocognitive disorder, a prognosis that seems adequate based on the diagnostic outcomes so far.

In this project each patient has given consent to use their medical record information for research purposes, information that is used to evaluate the clinical relevance of recorded behavioural signals in the interview situation, and that will be used for development and refinement of algorithms to optimise prognostic validity of our system. Ethical approval for the study was obtained from the Stockholm Ethical Board in decision dnr. 2018/1962-31.

Recordings During the Clinical Assessment

Each patient who fulfils the criteria for participation receives written information beforehand about the study, along with the summons for the examination. A week later a nurse calls the patient to ask if they want to participate in the study. After arrival to the clinic, the patient is asked again if they are still willing to participate and, if so, to sign the written consent form. The wristband is mounted and calibrated and the patient then walks with a physician to the examination room. Once the patient is seated, the eye tracker on the lower part of the tablet is calibrated. The researchers then leave the room and the multimodal recording starts. One technician continually monitors the recording a screen outside the room, as described in more detail below. The recording is terminated when the physical examination part starts, usually after 45–60 min of interviewing and testing. The examination is performed according to the normal clinical procedure at the clinic, but with some adaptations and additions to fit the purpose of our study: The first part of the interview is about the patient's background; living conditions, current and previous occupations, family situation, interests, memory problems or other cognitive problems, changes in personality, medication, sleep, medical history, and orientation in time and space (date, day of week, the location they are in). This part can be described as a conversation between the physician and the patient, and was carried out according to normal routine.

The second part includes a number of tasks that the patient performs to evaluate cognitive status, including the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005). This screening instrument includes various tasks to test performance in different cognitive domains, including drawing a line between letters and numbers (trail making), copying a figure, naming animals, drawing a clock that shows a certain time, immediate and delayed recall of words, generation of words, backwards counting, finger tapping, and abstract thinking. The figure copying and clock tests in particular are made to measure visuospatial constructional abilities and executive functioning (Charernboon, 2017). MoCA is a standard part of the examination protocol at the clinic, but for the patients who participate in the study it is adapted to be performed on a tablet, thereby allowing for detailed registration of pen movements and eye movements while the tasks are being performed, including trail making, the clock test, figure copying, and presentation of animals that the patient should name. For the tasks that involve drawing on the tablet, these drawings are mirrored in real time on a separate screen that the physician can see. The Boston Cookie Theft test (Giles et al., 1996) was added to the protocol for the purpose of this study, but is commonly used for screening. In this task the patient is asked to describe what is happening in a picture, a kitchen scene with a woman and two children. This picture is also shown on the tablet, allowing to sync eye movements and pupil changes with audio and video. When this part of the examination is over, the recording stops, and the wrist band is removed.

Export of Data

An export tool chain was created to export all of the files collected during the session in a standardised way, producing a set of CSV files. This step was performed by the clinician. The data was then stored on small hard drives in safety vaults. The data from the computers and phones was removed.

Further Tests and Collection of Other Biomarkers

After this first interview and examination of the patient, further data are collected to evaluate the cognitive status during the same and consecutive days, including more advanced cognitive testing, evaluation of mood and depressive symptoms, blood sample analysis, brain imaging (MRI, sometimes with the addition of PET if needed), and collection of CSF for analysis of biomarkers (levels of β-amyloid (Ab 40 and 42), tau, p-tau, and neurofilaments). The diagnostic decision is normally made within a week from the first interview, supported by the Combinosticsâ“¢ (Bruun et al., 2019) AI tool to combine results from the different sources of clinical information.

DATA ANALYSIS

In order to verify the validity of the data collected to date and to be able to compare against available measurements from each of the recorded patients, we perform a series of analyses and extract several descriptive physiological and behavioural metrics based on our captured modalities with a potential to serve as *digital biomarkers*. The extracted measures are summarised in **Table 3**. In most cases these metrics are calculated using basic statistics directly or indirectly over the collected data streams. For each of the extracted markers, we then calculate the correlation against a subset of *clinical assessment metrics* and biomarkers available as part of the regular memory clinic examination procedure. These are indicated in **Table 4**. A high correlation between one of our metrics and a clinical assessment variable indicates a potential suitability for that metric as a digital biomarker for AD. Below we describe how we extracted and analysed the various metrics from the captured modalities. As there is a large number of possible analysis that can be made, some have not been analysed in the scope of this work, and are instead suggestions for what can be analysed in the future. The modalities that were not analysed in this work were heart rate, skin conductivity, hand motion, and video. The others are described below.

Facial Gestures

The blendshape face data, including information on head motion and gaze, was captured from the "patient camera" sensor. From this data the following low level statistics were extracted: *smile mean*, *smile stdev*, *eyebrow stdev*, *head yaw/pitch/roll stdev*, *vertical/horizontal gaze shifts stdev* and *vertical/horizontal gaze shifts absolute mean*.

In addition we calculated the *correlation between vertical gaze shifts and vertical head movement* as well as the *correlation between horizontal gaze shifts and horizontal head movement*.

Gaze

From the gaze data we extracted the following digital biomarkers: *number of fixations*, *mean fixation duration*, *number of reading fixations*, *number of reading backtracks* (how many times during reading a fixation occurs to the left and above the previous fixation) and *percentage of reading backtracks*.

Language

The patient-clinician pairs of audio files were transcribed using Google Cloud Speech To Text in Swedish. The transcribed text was available as words with a start and end time and a confidence score for the translations. The transcribed patient text was used for language analysis. We extracted the following high-level metrics from the transcriptions: *Total number of words* and *total number of utterances* (during interview), *Average turn length* (Average number of words in a passage of patient speech with no in-between clinician speech) and *Percentage unique words* (number of unique words divided by total number of words). The ASR output was POS tagged with Universal-Dependencies formalism using the Stanford-NLP python package. These were used to develop 35 language features related to word type, open or closeness of word categories and average for all word categories. Examples of features are *Relative occurrence of adjectives, adverbials, verbs and nouns*.

Pen

The pen data from the parts of the clinical assessment where the patient was expected to draw something on the tablet was used to extract several different metrics, both independently for each part of the three drawing exercises in the MOCA test (trail, cube and clock) and for all of them taken together. The following metrics were calculated: *number of gaps* (how many times pen was lifted), *gap length, mean and standard deviation* (for how long was pen lifted), *drawing speed, mean and standard deviation* (how fast was the pen moving) and *pen pressure, mean and standard deviation*.

TABLE 3 | A demographic table with age, gender and education level for participants based on diagnostic group.

Demographic variable	Healthy	MCI	Alzheimer
Diagnostic group	14 (56%)	7 (28%)	4 (16%)
Age	60 (avg)	64.57 (avg)	64 (avg)
	3.39 (std)	4.11 (std)	3.9 (std)
Gender	11 females (78.5%)	1 female (14%)	4 females (100%)
	3 males (21.5%)	6 males (86%)	0 males (0%)
Education level	15.07 (avg)	14.14 (avg)	10.25 (avg)
	3.25 (std)	3.57 (std)	1.5 (std)

Pupil Dilation

From the gaze sensor data, we extracted pupil dilation measurements recorded together with the gaze tracking data, in order to study at pupil diameter across the each sessions. Measurements for left and right pupil were averaged, and rate-of-change was calculated by taking the difference between each consecutive reading. A median filter of length nine was applied to the rate-of-change signal to remove outliers due to sensor noise. We then extracted following metrics: *pupil maximum positive rate-of-change* (how fast can the pupil expand) and *pupil maximum negative rate-of-change* (how fast can the pupil contract), *pupil maximum rate-of-change* (how fast can the pupil change, regardless of direction), *pupil mean absolute rate-of-change* (how fast does pupil change on average) as well as *pupil diameter standard deviation*. All metrics were extracted independently for each of the exercises on the patient interface.

Thermal Emission

The "Patient camera (thermal)" sensor produces a thermal video, a thermal data file with temperatures given in Kelvin, and a RGB reference video. The RGB reference video is aligned to match the thermal video and thermal data file. Images from the RGB reference video and thermal video were extracted at one frame per second. Using the RGB reference frames it was then possible to apply the openpose pose extraction framework, Cao et al. (2021), to extract the pose of the patient. This was then used to determine a bounding box around the head, and the 10 highest values were then extracted from the corresponding region in the thermal images. The values were then aggregated and averaged for each minute of the interaction, and converted into percentages. Given the sequences of temperature readings with one value per minute, we extracted four metrics: *temp mean*, *temp stdev*, *temp rate-of-change mean* and *temp rate-of-change stdev*.

TABLE 4 | Summary table of clinical assessment metrics available, and an indication of which ones are used in the correlation analysis. From the MRI we have relative volume measurements for 248 brain regions; the table lists regions whose absolute Pearson correlation with diagnosis exceeds 0.7.

Modality	Assessment	Part of preliminary analysis
Medical assessment	Diagnosis	✓
	Moca-mis	✓
	MOCA	✗
	PHQ9	✗
	Background variables	✗
Spinal tap	Phosphorus tau	✓
	Ab42	✓
	Tau	✗
	Ab42Ab40	✗
	Ab42Ptau	✗
	NFL	✗
Neuropsychological tests	MMSE	✗
	RAVLT delayed recall	✗
	Rey complex figure	✗
	WAIS digit Symbolâ€“Coding	✗
MRI	Hippocampus total volume	✓
	Hippocampus (left, right)	✗
	Lateral ventricle (left, right)	✗
	Cerebellar vermal lobules (left, right)	✗
	Cerebrospinal fluid	✗
	Medial temporal lobe atrophy (left, right)	✗
	Cerebral cortex left GCA	✗
	Frontal lobe (left, right) GCA	✗
	Temporal lobe left GCA	✗
	Parietal lobe left GCA	✗

Voice

The recordings from the Microphone Array were split into patient and clinician audio files based on the angle of the sound source as reported by the microphone. The patient audio was used for voice analysis. In this preliminary analysis, minor irregularities were present in the voice splitting due to inaccuracy of direction of arrival (DoA) estimation, resulting in small segments of patient audio being labelled as clinician audio and vice versa, in particular in sections where there are overlapping speech (typically quite rate). More accurate methods can be applied by combining the four raw mic signals from the mic array.

Pauses and Speech Rate

All gaps in the patient's speech of a duration longer than 200 ms, with no intermediate speech from the clinician, were regarded as pauses. Start and end times for each word were retrieved from the output of the automatic speech recognition. We extracted several pause related metrics, such as *pause count* (total number of pauses), *average pause length* as well as *percentage pauses that are longer than 1, two or 3 s*. Furthermore, we extracted *speech rate* in syllables/second by counting number of syllables (approximated by number of vowels in the transcription) and divided by the total speech time.

Voice Quality Measures

In order to quantify vocal strength and breathiness, we calculated several acoustic measures of voice quality. All of the measures below are based on the relative amplitudes of the harmonics of the voice, where h1, h2 and h3 refers to the amplitude (in dB) of the first three harmonics, respectively, and a1, a2 and a3 denote the amplitude of the harmonic closest to the peak of the first, second and third formant, respectively. We extracted five metrics: h1h2 (h2–h1), h1h3, h1h3, h1a1, h1a2 and h1a3. We used REAPER[8] to extract fundamental frequency from all patient speech and SNACK[9] to extract formant trajectories. We measured the amplitudes of the harmonics in corresponding STFT spectrograms extracted using librosa[10] in Python. All measures were averaged over all voiced frames in the recording.

PRELIMINARY FINDINGS

In this section, we give some example analyses that illustrate how the digital biomarkers in the previous section may be connected to other diagnostic criteria. As our data gathering is far from complete, it is not possible to draw reliable conclusions about the diagnostic relevance from the material available thus far. Consequently, the analysis and results presented here are highly preliminary, and primarily serve to sketch the processes by which the digital biomarkers may be validated against other data available through the study. We deliberately omit *p*-values from the analyses so that readers are not tempted to treat the example analysis findings as statistically or scientifically significant.

At the time of writing 25 of 100 patients have been recorded. Our patients had a mean age of 61.92 years in the range 58–70 (standard deviation (4.16). 16 were females (64%) and 9 males (36%). Average length of education in years was 14.5 (standard deviation 3.55). From the 25 patients 4 patients were diagnosed with Alzheimer's disease, 7 with mild cognitive impairment and 14 received a diagnosis of subjective cognitive impairment, meaning the clinical examination found no clinical signs of impairment. Further demographic data is shown in **Table 3**.

Below we report how our extracted behavioural and physiological measures correlate to the following five biological biomarkers and clinical diagnostic measures:

These measures were chosen since they are relatively independent variables within our dataset with a strong correlation to AD diagnosis (Moca-MIS 0.70, p-tau 0,65, Ab42 -0.647, Hippocampus, -0.766).

Moca Memory Index Score (MoCA-MIS) is a sub-scoring of MOCA that focus on memory tasks. The MoCA-MIS is calculated by adding the number of words remembered in free delayed recall, category-cued recall, and multiple choice–cued recall multiplied by 3, 2 and 1, respectively, with a score ranging from 0 to 15 Julayanont et al. (2014). MOCA-MIS was chosen over full scale MOCA since it has a stronger correlation to diagnostic then the full MOCA test. Ab42 and p-tau are both linked to AD pathology. The scientific debate regarding the relationship and validity of Ab42 and p-tau as diagnostic criteria in AD is ongoing. We chose to present Ab42 and p-tau independently although they have good diagnostic validity as a single biomarker in our dataset (Ab42/p-tau, −0.7179). Hippocampus was chosen since it is a well studied brain region closely tied to AD pathology. In our preliminary analysis of the data collected to date, we found many correlations between our extracted metrics and the above measures (please see 8). Below we report the most prominent ones (**Figure 6**). We used Pearson correlations for all our correlation measurements. We made a comparison between Pearson and Spearman correlations but no major differences were found (mean average difference −0.01 ± 0.17). In our current situation, where the amount of data is very small, we believe that making distributional assumptions (i.e., the Pearson correlation) offers the most appropriate bias-variance trade-off, especially since the analysis is only intended to be preliminary.

Facial Gestures

We found that the Moca-MIS score correlated negatively with *smile mean* (-0.62) and *smile standard deviation* (-0.68). For the gaze data captured by the iphone during the interview part, we found a negative correlation of *horizontal gaze* (sideways gaze movements) and diagnosis of -0.54 for *horizontal gaze absolute mean* and -0.5 for *horizontal gaze standard deviation*. These statistics also correlated positively with hippocampus total volume (0.57 and 0.54 respectively).

Gaze

From the data captured by the gaze tracker during interactions with the ipad, we found that the total number

[8] https://github.com/google/REAPER

[9] http://www.speech.kth.se/snack/

[10] https://librosa.org/doc/latest/index.html

of fixations correlated with diagnosis (−0.32) and with hippocampus total volume (0.67). Further, mean fixation duration correlated with diagnosis (0.45) and hippocampus total volume (−0.78).

Language

Total word count correlated with Moca-MIS (0.36), Ab42 (0.51) hippocampus total volume (0.45), while *Percentage unique words* correlated with Moca-MIS (0.37), Ab42 (0.54) and hippocampus total volume (0.44). For the word type metrics, *relative occurance of Adjectives* was the most relevant feature with a correlation with Moca-MIS (0.44), Ab42 (0.61) hippocampus total volume (0.54).

Pupil dilation

The metric *pupil maximal absolute rate-of-change* generally correlated well with several of the biomarkers, but correlations varied across the different sub tasks. Highest correlations was achieved for tasks that involved drawing (path, cube and clock tests): for clock drawing test and cube test, correlation with diagnosis was −0.47 and −0.56 respectively, Moca-MIS (0.6 and 0.54), p-tau (0.8 and 0.75) and Ab42 (0.9 and 0.77).

Thermal Emissions

For face temperature measurements captured with the “Patient camera (thermal)” sensor we found that *temp mean* correlated with diagnosis (−0.41) and hippocampus total volume (0.65) while *temp rate-of-change mean* correlated with diagnosis (0.37) and hippocampus total volume (−0.63).

Pen Motion and Pressure

Figures 7 and **8** show typical output from two of the drawing tasks for sample subjects of each of the diagnosis categories. Looking at the statistics of pen motion and pen pressure, we found that two features were particularly interesting: *mean drawing gap length* correlated with diagnosis (0.62), Moca-MIS (−0.61) and Hippocampus total volume (−0.58), and *mean pen pressure* correlated with p-tau (−0.88) and Hippocampus total volume (0.86).

Voice

Two classes of voice related features are included in this analysis: voice source metrics and pause/speech rate features. Several of the extracted voice quality metrics (breathiness/vocal strength) showed correlation to diagnosis and biomarkers. The most relevant were *h1h3* that correlated with diagnosis (0.68) and

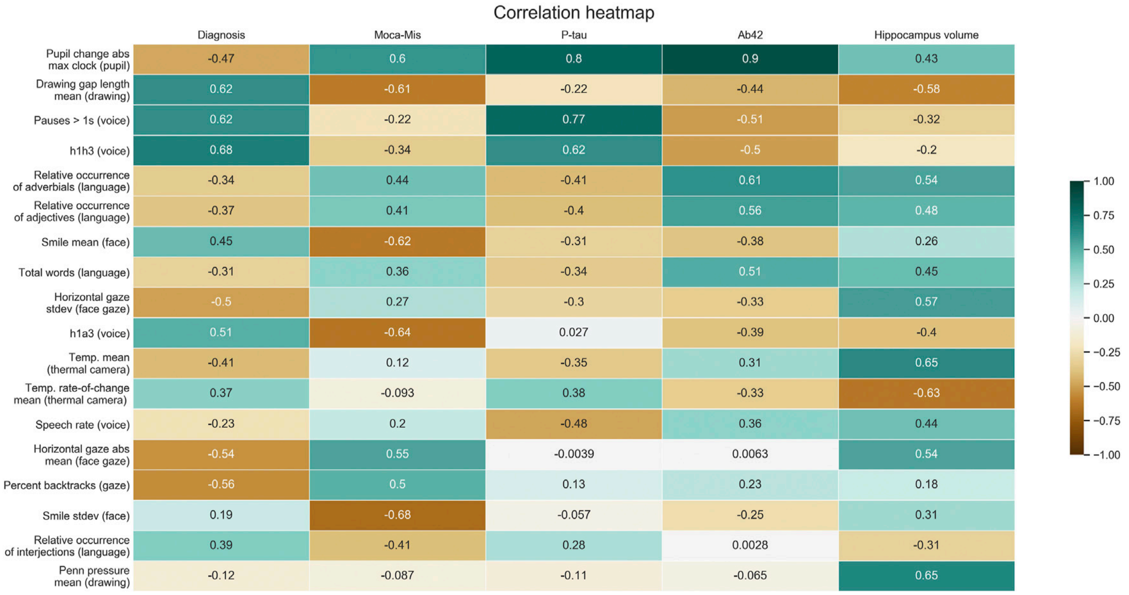


FIGURE 6 | Summary of correlations between selected digital biomarker candidate metrics and clinical assessment measures.

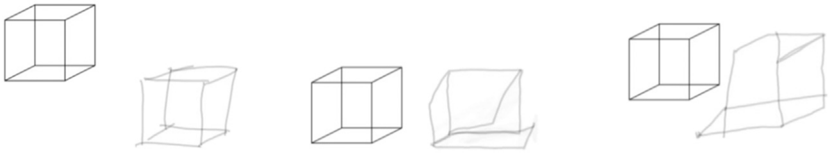


FIGURE 7 | Cube drawing based on category. From left to right: Healthy, MCI, Alzheimer.

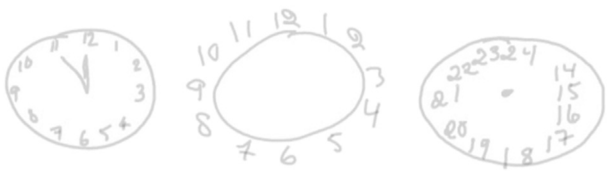


FIGURE 8 | Clock drawing based on category. From left to right: Healthy, MCI, Alzheimer.

p-tau (0.62) and *h1a3* that correlated with diagnosis (0,51) and Moca-MIS (−0.64). *Percentage pauses longer than 1 s* correlated with diagnosis (0.62) and p-tau (0.77) while *speech rate* correlated with p-tau (−0.48) and hippocampus total volume (0.44).

DISCUSSION

Our study describes how to design and implement a multimodal sensor recording system in a clinical setting. Furthermore we report our preliminary findings from our sensor data capture. Several of the digital biomarkers abstracted from sensor data were highly correlated to both the diagnostic outcome and to biomarkers of Alzheimer's disease, suggesting that a multimodal approach has the potential to complement and improve current diagnostic processes. In the remainder of this section, we discuss the results of the preliminary analysis of the digital biomarkers we studied, and consider the implications of our data capture and its findings for dementia detection and treatment.

Discussion of Analysis Findings

For the purposes of this article, a digital biomarker is useful if it is sensitive to early signs of AD, or informative about the current stage of the patient's disorder, or both. At present, three biomarkers are considered to be central for a state-of-the-art evaluation of a possible neurocognitive disorder:

- levels of β-amyloid (levels of Ab 42, and/or the ratio between Ab42/Ab40);
- levels of Tau (Both Total Tau and P-tau); and
- cerebral atrophy (including both in specific regions, such as the entorhinal region and hippocampus, and general atrophy (including enlarged ventricles).

A high-quality and detailed examination will include all three biomarkers, and their coexistence, which was performed for all patients included in our study (along with other in-depth assessments, as described earlier). Due to costs, limited resources, and the invasive nature of these measurements, it is important to identify for which patients this extensive examination is needed and for which patients it is not. It is obviously advantageous if this can be done in a non-invasive and non-intrusive way. With the assumption that the above biomarkers in combination adequately reflect the underlying neuropathology with a high level of sensitivity and specificity, digital biomarkers of clinical utility will need to demonstrate a high correlation with these existing biomarkers.

Our data analysis covered both established and novel digital biomarkers. For the former, our findings were in line with previous AD research. *Pause length* and vocal strength metrics *h1h3*, specifically, correlated with AD diagnosis, β amyloid-42 protein, and p-tau. Overall, we also found that voice measures correlated more strongly with clinical assessment metrics than language measures did. Voice features may generally be more useful than language measures for early dementia detection, since the semantic features of language are more obviously disrupted in the later stages of AD. As our dataset contains only 3 individuals diagnosed with AD, our findings are likely more informative for indicating utility in early diagnostics, than for the ability of different biomarkers to distinguish AD patients from the two less-affected patient groups we considered.

Another promising digital biomarker we studied that has been previously proposed for AD assessment was pupil change. We found that maximum change during cognitively taxing tasks strongly correlated with both diagnosis, moca-mis, p-tau, Ab42, and hippocampal volume. The fact that a difference was noticeable between non-taxing (cookie test) and taxing (clock, cube, path drawing) tasks shows that this might potentially be a useful biomarker in combination with a cognitive test. Unlike voice and language, this digital biomarker quantifies physiological responses in the patient that clinicians cannot feasibly detect, which increases its potential to complement existing diagnostic procedures.

We also identified several promising new digital biomarkers. In particular, the mean head temperature rate of change correlated strongly with diagnosis, p-tau, Ab42, and hippocampal volume. The pen-drawing gap length correlated strongly with diagnosis, moca-mis, Ab42 and hippocampus. Furthermore it was highly correlated to vocal pause length measurements (correlation coefficient 0.72). Both pause length and pen-drawing gap length are likely related to sympathetic nervous system responses, which differ for patients with AD or MCI, compared to those with no objective impairment (Borson et al., 1989). This potential utility in early detection can be contrasted against assessments of the drawings themselves, where only 53.3% of normal elderly can copy the cube correctly, although most are able to correctly draw the clock (Charernboon, 2017). Without pen data, drawing tests in general are thus sensitive detectors of AD but not MCI.

Tasks and Sensors

When considering different digital biomarkers and their capture, it is worth distinguishing between task-dependent and task-independent digital biomarkers. A task-independent digital biomarker is one that can be gathered at any (or all) point in the interaction. As such, these are arguably more valuable since they are much easier to capture, and do not put constraints on the specifics of the clinical interview. Among the different measures in our study, voice and language features can be seen as mostly task-independent while quantities extracted from gaze, pupil, and drawing depend on a task. Although task-dependent digital biomarkers are more specific and targeted, which might increase accuracy and specificity, that has to be weighted against the relative increase in complexity of the associated data capture. A microphone can simply record a person's voice while gaze, pupil and drawing sensors all depend on a well-designed task for gathering data that enables accurate diagnosis.

All things considered, microphones are arguably the most useful among those we considered for dementia detection and diagnostics. The relative ease of unobtrusive audio capture and the ability to extract powerful features (e.g., pause length, voice source h1a3) makes it a cheap and useful diagnostic tool. Furthermore, automatic transcripts of the gathered interview audio can also be used to extract linguistic digital biomarkers via text processing, although this may be less relevant for early diagnosis and the digital tools and their maturity will differ across languages, whereas the tools used to extract voice measures do not.

Because of the notable correlation of pupilary data with AD diagnosis, p-tau, Ab42, and hippocampal volume, device-mounted eye-trackers capable of accurately measuring pupil size also have shown potential for augmenting and improving diagnostic procedures, and there might be promise in building an application that combines pupilary measurements with a cognitive test to build more accurate automatic screening tests for dementia. Measures based on drawing and pen pressure have the drawback that they mainly appeared useful for diagnosing between healthy control and AD, a result that should be interpreted with caution since only three individuals with AD were included in the preliminary analysis. That said, various associated digital biomarkers such as gap length show potential and merit more study.

Broader Implications

The non-invasive and non-intrusive nature of our data-capture setup brings several benefits. Non-invasive procedures generally have lower cost and complexity than invasive ones, and also limit the need for health-care personnel since the risk of adverse effects and reactions is much lower. Our non-intrusive data capture does not alter the diagnostic interview in a meaningful way. This is helpful both for obtaining ecologically valid data and in building trust for data-driven diagnostics among both clinicians and patients. By basing the data gathering on affordable and widely-available consumer electronics we hope to demonstrate how to the access to sensor-based diagnostic tools for dementia detection and monitoring can be democratised.

A key strength of using a multimodal approach as described in this article is that the different measurements can reinforce each others' predictive power while limiting risks from data loss and inaccuracies in the data pipeline. Our in-depth descriptions of our technical setup, data capture procedure, and data processing should enable independent replication of our findings using similar sensors. To further simplify such replication, we will release the the code used for the data capture and processing as open source.

An important consideration in the bigger picture is the temporal and neuronal aspect of AD. Although the diagnostic criteria is limited to healthy, MCI or AD, beneath the diagnosis lies a progressive disorder with a unique pattern of brain functioning for each patient. Assessment of AD is an assessment of the individual's cognitive functions and their deficits. Streamlined diagnostics offer the potential of continuous assessment of cognitive functions for individuals in the MCI/AD group. For patients with MCI, deficits are specific to certain areas of functioning and continuous assessment enables adaptive care with limited restrictions. This is likely to improve the daily life of the patient, which in turn might help the patient not progress to AD (through better quality of life and reduced life stressors). Continuous screening as part of behavioural interventions might help furthermore develop a virtuous cycle of improved understanding of the disorder, through data capture that leads to better targeted interventions.

If non-invasive measurements can accurately predict underlying brain atrophy in different areas, that also opens the door to a future where quick tests can quantify disease progress. This could help in the quest to find a cure, since behavioural interventions and targeted pharmaceutical drugs might be used to target specific brain atrophies caused by the disorder.

CONCLUSION

We have described a non-invasive and non-intrusive system for collecting synchronous behavioural and physiological data in order to facilitate detection of early signs of Alzheimer's disease, based on a large and diverse set of modalities including speech, gaze, pupillometry, facial motion capture, drawing, heart rate and thermal data in existing clinical assessments of dementia, and also used the initial data thus gathered for a preliminary analysis of selected digital biomarkers available through our approach, and their diagnostic value.

The modalities we capture allow both behavioural and physiological measurements in an objective and quantitative manner, and thus complementing the intuitive and qualitative observations made by the assessing clinician. The studied modalities may not only quantify the observations and "gut feeling" of the clinician, but can also measure aspects of the patient and interaction that are inaccessible to human perception. Our work demonstrates that the proposed approach is feasible with commodity hardware and open-source software that we are preparing for public release.

Our multimodal approach to digital biomarkers has the potential to improve precision in patient selection for further and more invasive examinations, thereby saving personnel-time and financial resources for society, and avoiding unnecessary delays, suffering, and discomfort for patients. While existing full-fledged diagnostic procedures are advanced, they still result in a troubling amount of misdiagnoses (Villemagne et al., 2018; Gauthreaux et al., 2020). To the extent that systems and measurements of the kind described in this article also can contribute to diagnostic accuracy, that should benefit patients and their families in several ways, including reducing exposure to unnecessary medication with negative side-effects and avoiding life-quality losses associated with a false positive diagnosis.

Our analysis finds that single modalities can be used for AD prediction in isolation. Some of these have not been reported previously: Our preliminary results indicate that head temperature change and drawing gap length are two new digital biomarkers that correlated with AD diagnosis and biological biomarkers. Pupillary response has been used for AD prediction but to our knowledge not in the context of cognitively demanding tasks. Other preliminary results confirm what is known from previous work, such as the correlation of pause length, vocal strength and gaze patterns with a dementia diagnosis. This demonstrates that a broad and inclusive data-gathering approach has the potential to discover new digital biomarkers of clinical utility, which in turn can serve as further clues to understand underlying mechanisms of AD and other neurocognitive disorders. The fact that isolated modalities correlate well with established biomarkers and the clinical diagnosis also suggests the potential of combining different modalities and measures for further improved diagnostic accuracy. It should be noted that all of the metrics explored in the current study are manually crafted features. As is well known from machine learning e.g. in speech and image processing, automatically learned features generally outperform hand crafted features when sufficient amounts of data are available. Machine learning based feature extraction, prediction and classification methods will be a central area of exploration as these data collection efforts continue.

As it stands, a limitation of the results presented in this paper is the relatively small number of patients, which does not allow statistically rigorous conclusions nor discriminating between different types of neurocognitive disorders. Our preliminary results therefore mainly pertain to patients with AD, the most common dementia diagnosis. Another limitations is that, also for reasons of statistical power, we have only focused on measures relevant to atrophy in brain regions known to be especially affected by AD. In future studies with more patients, we intend to explore measures and modalities that associate with changes in a broader range of brain regions.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because Since clinical patient data is used, the access to data is highly restricted but we will work with frontiers to comply with frontiers guidelines as well as Swedish regulatory guidelines related to clinical patient data. Requests to access the datasets should be directed to Jonas Beskow beskow@kth.se.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Regionala Etikprövningsnämnden i Stockholm. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PJ Planning, design, and setup of the data collection system. Responsible for the majority of the implementation of the data collection system at Karolinska University Hospital (KUH). Manuscript writing and editing, primarily sections 4–5. Data analysis. BM Setup of clinical data capture KUH. Data analysis and feature extraction. Manuscript writing and editing, mainly sections 5–8. KH Setup of clinical data capture KUH. Manuscript writing, primarily the medical aspects of the manuscript. GH Setup of clinical data capture KUH. Manuscript writing and editing. TK Technical and manuscript discussions, and data quality assurance. OM Technical discussions and data quality assurance. GH Clinical assessment KUH JH Clinical data capture KUH MK Co-PI of the project. Overall strategic planning and discussions on clinical data collection activities. HK PI of the project. Planning and disscussions on technical activities. Manuscript editing. JG Co-PI of the project. He has been in charge of section 3 Related work in the current paper. JB Co-PI of the project. Coordination and planning of data collection and data analysis. Speech and multimodal data processing. Manuscript writing, primarily sections 5–6.

FUNDING

This research was funded by Swedish Foundation of Strategic Research (SSF), project EACare under Grant No RIT15-0107.

REFERENCES

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* 136, 3727–3737. doi:10.1093/brain/awt269

Algotsson, A., Viitanen, M., Winblad, B., and Solders, G. (1995). Autonomic dysfunction in Alzheimer's disease. *Acta Neurolo. Scand.* 91, 14–18. doi:10.1111/j.1600-0404.1995.tb05836.x

Allan, L. M., Kerr, S. R. J., Ballard, C. G., Allen, J., Murray, A., McLaren, A. T., et al. (2005). Autonomic function assessed by heart rate variability is normal in alzheimer's disease and vascular dementia. *Dement Geriatr. Cogn. Disord.* 19, 140–144. doi:10.1159/000082885

Anzengruber, B., and Riener, A. (2012). "FaceLight - potentials and drawbacks of thermal imaging to infer driver stress," in ACM SIGCHI international conference on automotive user Interfaces and interactive vehicular applications, Portsmouth, NH, October 2012, 209–216. doi:10.1145/2390256.2390292

Asgari, M., Kaye, J., and Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's Demen. Transl. Res. Clin. Intervent.* 3, 219–228. doi:10.1016/j.trci.2017.01.006

Asplund, K., Norberg, A., Adolfsson, R., and Waxman, H. M. (1991). Facial expressions in severely demented patients?a stimulus-response study of four patients with dementia of the Alzheimer type. *Int. J. Geriat. Psychiatry* 6, 599–606. doi:10.1002/gps.930060809

Beltrán, J., García-Vázquez, M. S., Benois-Pineau, J., Gutierrez-Robledo, L. M., and Dartigues, J.-F. (2018). Computational techniques for eye movements analysis towards supporting early diagnosis of Alzheimer's disease: a review. *Comput. Math. Methods Med.* 14, 2676409. doi:10.1155/2018/2676409

Blennow, K., de Leon, M. J., and Zetterberg, H. (2006). Alzheimer's disease. *Lancet* 368, 387–403. doi:10.1016/s0140-6736(06)69113-7

Borson, S., Barnes, R. F., Veith, R. C., Halter, J. B., and Raskind, M. A. (1989). Impaired sympathetic nervous system response to cognitive effort in early Alzheimer's disease. *J. Gerontol.* 44, M8–M12. doi:10.1093/geronj/44.1.m8

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., and Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: a review. *Front. Psychol.* 8, 269. doi:10.3389/fpsyg.2017.00269

Bruun, M., Frederiksen, K. S., Rhodius-Meester, H. F. M., Baroni, M., Gjerum, L., Koikkalainen, J., et al. (2019). Impact of a clinical decision support tool on dementia diagnostics in memory clinics: the PredictND validation study. *Curr. Alzheimer. Res.* 16, 91–101. doi:10.2174/1567205016666190103152425

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology* 14, 71–91. doi:10.1080/026870300401603

Burton, K. W., and Kaszniak, A. W. (2006). Emotional experience and facial expression in Alzheimer's disease. *Aging Neuropsychol. Cogn.* 13, 636–651. doi:10.1080/13825580600735085

Cai, Y., Li, L., Xu, C., and Wang, Z. (2020). The effectiveness of non-pharmacological interventions on apathy in patients with dementia: a systematic review of systematic reviews. *Worldviews Evidence-Based Nurs.* 17, 311–318. doi:10.1111/wvn.12459

Calzà, L., Gagliardi, G., Favretti, R. R., and Tamburini, F. (2020). Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Comp. Speech Lang.* 65, 101113. doi:10.1016/j.csl.2020.101113

Cao, Q., Tan, C.-C., Xu, W., Hu, H., Cao, X.-P., Dong, Q., et al. (2020). The prevalence of dementia: a systematic review and meta-analysis. *J. Alzheimers Dis.* 73, 1157–1166. doi:10.3233/JAD-191092

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2021). Openpose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 172–186. doi:10.1109/tpami.2019.2929257

Castellani, R. J., Lee, H.-g., Siedlak, S. L., Nunomura, A., Hayashi, T., Nakamura, M., et al. (2009). Reexamining Alzheimer's disease: evidence for a protective role for amyloid-β protein precursor and amyloid-β. *J Alzheimers Dis.* 18, 447–452. doi:10.3233/jad-2009-1151

Charernboon, T. (2017). Diagnostic accuracy of the overlapping infinity loops, wire cube, and clock drawing tests for cognitive impairment in mild cognitive impairment and dementia. *Int. J. Alzheimer's Dis.* 2017, 5289239. doi:10.1155/2017/5289239

Chen, R., Jankovic, F., Marinsek, N., Foschini, L., Kourtis, L., Signorini, A., et al. (2019). "Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams," in ACM SIGKDD international Conference on knowledge discovery & data mining, Anchorage, AK, United States, July 2019 (Association for Computing Machinery), 2145–2155. doi:10.1145/3292500.3330690

Cho, Y. (2018). "Automated mental stress recognition through mobile thermal imaging," in International Conference on affective Computing and intelligent interaction, San Antonio, TX, United States, Oct. 2017 (IEEE). doi:10.1109/ACII.2017.8273662

Chung, P.-C., Hsu, Y.-L., Wang, C.-Y., Lin, C.-W., Wang, J.-S., and Pai, M.-C. (2012). "Gait analysis for patients with Alzheimer's disease using a triaxial accelerometer," in IEEE international Symposium on Circuits and systems, Seoul, Korea (South), May 2012 (IEEE).

da Silva, V. P., Ramalho Oliveira, B. R., Tavares Mello, R. G., Moraes, H., Deslandes, A. C., and Laks, J. (2018). Heart rate variability indexes in dementia: a systematic review with a quantitative analysis. *Curr. Alzheimer Res.* 15, 80–88. doi:10.2174/1567205014666170531082352

de la Fuente Garcia, S. Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J Alzheimers Dis.* 78, 1547–1574. doi:10.1007/s10286-009-0035-

de Vilhena Toledo, M. A., and Junqueira, L. F. (2010). Cardiac autonomic modulation and cognitive status in Alzheimer's disease. *Clin. Auton. Res.* 20, 11–17. doi:10.1007/s10286-009-0035-0

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2015). The Geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Trans. Affective Comput.* 7, 190–202. doi:10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). "Recent developments in opensmile, the Munich open-source multimedia feature extractor," in ACM international Conference on multimedia, Barcelona, Spain, October 2013 (Association for Computing Machinery), 835–838. doi:10.1145/2502081.2502224

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the Munich versatile and fast open-source audio feature extractor," in ACM international Conference on multimedia, Firenze Italy, October 2010 (Association for Computing Machinery), 1459–1462. doi:10.1145/1873951.1874246

Ferreira, D., Verhagen, C., Hernández-Cabrera, J. A., Cavallin, L., Guo, C.-J., Ekman, U., et al. (2017). Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Scientific Rep.* 7, 46263. doi:10.1038/srep46263

Ferri, C. P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., et al. (2005). Global prevalence of dementia: a delphi consensus study. *Lancet* 366, 2112–2117. doi:10.1016/s0140-6736(05)67889-0

Garbutt, S., Matlin, A., Hellmuth, J., Schenk, A. K., Johnson, J. K., Rosen, H., et al. (2008). Oculomotor function in frontotemporal lobar degeneration, related disorders and Alzheimer's disease. *Brain* 131, 1268–1281. doi:10.1093/brain/awn047

Garre-Olmo, J., Faúndez-Zanuy, M., López-de Ipiña, K., Calvó-Perxas, L., and Turró-Garriga, O. (2017). Kinematic and pressure features of handwriting and drawing: preliminary results between patients with mild cognitive impairment, alzheimer disease and healthy controls. *Curr. Alzheimer Res.* 14, 960–968. doi:10.2174/1567205014666170309120708

Gatouillat, A., Dumortier, A., Perera, S., Badr, Y., Gehin, C., and Sejdić, E. (2017). Analysis of the pen pressure and grip force signal during basic drawing tasks: the timing and speed changes impact drawing characteristics. *Comput. Biol. Med.* 87, 124–131. doi:10.1016/j.compbiomed.2017.05.020

Gauthreaux, K., Bonnett, T. A., Besser, L. M., Brenowitz, W. D., Teylan, M., Mock, C., et al. (2020). Concordance of clinical Alzheimer diagnosis and neuropathological features at autopsy. *J. Neuropathol. Exp. Neurol.* 79, 465–473. doi:10.1093/jnen/nlaa014

Gavas, R., Chatterjee, D., and Sinha, A. (2017). "Estimation of cognitive load based on the pupil size dilation," in IEEE International Conference on Systems, Man, and Cybernetics (SMC). Banff, AB, October 5–8, 2017 (IEEE). doi:10.1109/smc.2017.8122826

Giles, E., Patterson, K., and Hodges, J. R. (1996). Performance on the Boston Cookie theft picture description task in patients with early dementia of the Alzheimer's type: missing information. *Aphasiology* 10, 395–408. doi:10.1080/02687039608248419

Granholm, E. L., Panizzon, M. S., Elman, J. A., Jak, A. J., Hauger, R. L., Bondi, M. W., et al. (2017). Pupillary responses as a biomarker of early risk for Alzheimer's disease. *Jad* 56, 1419–1428. doi:10.3233/jad-161078

Hafiz, P., and Bardram, J. E. (2020). The ubiquitous cognitive assessment tool for smartwatches: design, implementation, and evaluation study. *JMIR Mhealth and Uhealth* 8, e17506. doi:10.2196/17506

Hafiz, P., Miskowiak, K. W., Kessing, L. V., Jespersen, A. E., Obenhausen, K., Gulyas, L., et al. (2019). The internet-based cognitive assessment tool: system design and feasibility study. *JMIR Formative Res.* 3, e13898. doi:10.2196/13898

Haider, F., de la Fuente, S., and Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE J. Selected Top. Signal Process.* 14, 272–281. doi:10.1109/JSTSP.2019.2955022

Håkansson, K., Ngandu, T., and Kivipelto, M. (2018). "The patient with cognitive impairment," in *Treatable and potentially preventable dementias.* Editor V. Hachinsky (Cambridge, United Kingdom: Cambridge University Press), 52–80.

Hamet, P., and Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism* 69, S36–S40. doi:10.1016/j.metabol.2017.01.011

Hardy, J., and Higgins, G. (1992). Alzheimer's disease: the amyloid cascade hypothesis. *Science* 256, 184–185. doi:10.1126/science.1566067

Henneges, C., Reed, C., Chen, Y.-F., Dell'Agnello, G., and Lebrec, J. (2016). Describing the sequence of cognitive decline in Alzheimer's disease patients: results from an observational study. *J Alzheimers Dis* 52, 1065–1080. doi:10.3233/jad-150852

Insel, P., Donohue, M., Berron, D., Hansson, O., and Mattsson-Carlgren, N. (2020). Time between milestone events in the Alzheimer's disease amyloid cascade. *bioRxiv*. doi:10.1101/2020.05.18.103226

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's Demen.* 14, 535–562. doi:10.1016/j.jalz.2018.02.018

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., et al. (2014). "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in Workshop on computational Linguistics and clinical psychology: from linguistic Signal to clinical reality, Baltimore, Maryland, June 2014 (Association for Computational Linguistics), 27–37. doi:10.1109/smc.2017.8122826

Jonell, P., Bystedt, M., Fallgren, P., Kontogiorgos, D., Lopes, J., Malisz, Z., et al. (2018). "FARMI: a framework for recording multi-modal interactions," in International Conference on language Resources and evaluation, Miyazaki, Japan

Julayanont, P., Brousseau, M., Chertkow, H., Phillips, N., and Nasreddine, Z. S. (2014). Montreal cognitive assessment memory index score (MoCA-MIS) as a predictor of conversion from mild cognitive impairment to Alzheimer's disease. *J. Am. Geriatr. Soc.* 62, 679–684. doi:10.1111/jgs.12742

Kepp, K. P. (2017). Ten challenges of the amyloid hypothesis of Alzheimer's disease. *J. Alzheimers Dis.* 55, 447–457. doi:10.3233/JAD-160550

Kivipelto, M., Mangialasche, F., and Ngandu, T. (2017). Can lifestyle changes prevent cognitive impairment? *Lancet Neurol.* 16, 338–339. doi:10.1016/s1474-4422(17)30080-7

Koikkalainen, J. R., Rhodius-Meester, H. F., Rhodius-Meester, H. F. M., Frederiksen, K. S., Bruun, M., Hasselbalch, S. G., et al. (2019). Automatically computed rating scales from mri for patients with cognitive disorders. *Eur. Radiol.* 29, 4937–4947. doi:10.1007/s00330-019-06067-1

Komeili, M., Pou-Prom, C., Liaqat, D., Fraser, K. C., Yancheva, M., and Rudzicz, F. (2019). Talk2me: automated linguistic data collection for personal assessment. *PLoS ONE* 14, e0212342. doi:10.1371/journal.pone.0212342

König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., and Robert, P. (2018). Fully automatic speech-based analysis of the semantic verbal fluency task. *Dement Geriatr. Cogn. Disord.* 45, 198–209. doi:10.1159/000487852

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 1, 112–124. doi:10.1016/j.dadm.2014.11.012

Kourtis, L. C., Regele, O. B., Wright, J. M., and Jones, G. B. (2019). Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *npj Digital Med.* 2, 1–9. doi:10.1038/s41746-019-0084-2

Kumar, D. K. V., Choi, S. H., Washicosky, K. J., Eimer, W. A., Tucker, S., Ghofrani, J., et al. (2016). Amyloid-β peptide protects against microbial infection in mouse and worm models of Alzheimer's disease. *Sci. Translational Med.* 8, 340ra72. doi:10.1126/scitranslmed.aaf1059

Lane, C. A., Hardy, J., and Schott, J. M. (2018). Alzheimer's disease. *Eur. J. Neurol.* 25, 59–70. doi:10.1111/ene.13439

Lang, L., Clifford, A., Wei, L., Zhang, D., Leung, D., Augustine, G., et al. (2017). Prevalence and determinants of undetected dementia in the community: a systematic literature review and a meta-analysis. *BMJ Open* 7, e011146, doi:10.1136/bmjopen-2016-011146

Lee, G., Nho, K., Kang, B., Sohn, K.-A., and Kim, D. (2019a). Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific Rep.* 9, 1–12. doi:10.1038/s41598-018-37769-z

Lee, J. E., Yang, S. W., Ju, Y. J., Ki, S. K., and Chun, K. H. (2019b). Sleep-disordered breathing and Alzheimer's disease: a nationwide cohort study. *Psychiatry Res.* 273, 624–630. doi:10.1016/j.psychres.2019.01.086

Lewis, G. F., Gatto, R. G., and Porges, S. W. (2011). A novel method for extracting respiration rate and relative tidal volume from infrared thermography. *Psychophysiol.* 48, 877–887. doi:10.1111/j.1469-8986.2010.01167.x

Li, H., Liu, C.-C., Zheng, H., and Huang, T. Y. (2018). Amyloid, tau, pathogen infection and antimicrobial protection in Alzheimer's disease – conformist, nonconformist, and realistic prospects for AD pathogenesis. *Translational neurodegeneration* 7, 34. doi:10.1186/s40035-018-0139-3

Long, J. M., and Holtzman, D. M. (2019). Alzheimer disease: an update on pathobiology and treatment strategies. *Cell* 179, 312–339. doi:10.1016/j.cell.2019.09.001

Malisz, Z., Jonell, P., and Beskow, J. (2019). "The visual prominence of whispered speech in Swedish," in International Congress of phonetic sciences

Matsushita, M., Yatabe, Y., Koyama, A., Katsuya, A., Ijichi, D., Miyagawa, Y., et al. (2018). Are saving appearance responses typical communication patterns in Alzheimer's disease? *PLOS ONE* 13, e0197468. doi:10.1371/journal.pone.0197468

McGirr, S., Venegas, C., and Swaminathan, A. (2020). Alzheimer's disease: a brief review. *J. Exp. Neurol.* 1, 89–98.

Migliaccio, R., Tanguy, D., Bouzigues, A., Sezer, I., Dubois, B., Le Ber, I., et al. (2020). Cognitive and behavioural inhibition deficits in neurodegenerative dementias. *Cortex* 131, 265–283. doi:10.1016/j.cortex.2020.08.001

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., et al. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *J Alzheimers Dis* 58, 373–387.doi:10.3233/jad-160507

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019). Dementia detection using automatic analysis of conversations. *Comp. Speech Lang.* 53, 65–79. doi:10.1016/j.csl.2018.07.006

Molitor, R. J., Ko, P. C., and Ally, B. A. (2015). Eye movements in Alzheimer's disease. *J Alzheimers Dis* 44, 1–12. doi:10.3233/jad-141173

Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. *J. Clin. Exp. Neuropsychol.* 40, 917–939. doi:10.1080/13803395.2018.1446513

Nam, U., Lee, K., Ko, H., Lee, J.-Y., and Lee, E. C. (2020). Analyzing facial and eye movements to screen for Alzheimer's disease. *Sensors* 20, 5349. doi:10.3390/s20185349

Nasreddine, Z. S., Phillips, N. A., Bã©dirian, V. r., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi:10.1111/j.1532-5415.2005.53221.x

Negami, M., Maruta, T., Takeda, C., Adachi, Y., and Yoshikawa, H. (2013). Sympathetic skin response and heart rate variability as diagnostic tools for the differential diagnosis of Lewy body dementia and Alzheimer's disease: a diagnostic test study. *BMJ Open* 3, e001796. doi:10.1136/bmjopen-2012-001796

Phillips, N., Al-Yawer, F., Giroud, N., Rehan, S., Pappadatos, Z., Mick, P., et al. (2020). "Sensory function, cognition, and brain structure in SCD, MCI, and AD: initial findings from the COMPASS-ND study," in Alzheimer's association international conference *Clinical Manifestations* 16, e044056. doi:10.1002/alz.044056

Pistono, A., Jucla, M., Barbeau, E. J., Saint-Aubert, L., Lemesle, B., Calvet, B., et al. (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *J Alzheimers Dis* 50, 687–698. doi:10.3233/jad-150408

Prince, M. J. (2015). World Alzheimer Report 2015. The global impact of dementia: an analysis of prevalence, incidence, cost and trends (Alzheimer's Disease International)

Reitan, R. M., and Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: theory and clinical interpretation*. London: Neuropsychology Press.

Roalf, D. R., Rupert, P., Mechanic-Hamilton, D., Brennan, L., Duda, J. E., Weintraub, D., et al. (2018). Quantitative assessment of finger tapping characteristics in mild cognitive impairment, Alzheimer's disease, and Parkinson's disease. *J. Neurol.* 265, 1365–1375. doi:10.1007/s00415-018-8841-8

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans. Audio Speech Lang. Process.* 19, 2081–2090. doi:10.1109/tasl.2011.2112351

Ruminski, J. and Kwasniewska, A. (2017). Evaluation of respiration rate using thermal imaging in mobile conditions. *Appl. Infrared Biomed. Sci.* 14, 311–346. doi:10.1007/978-981-10-3147-2_18

Scarmeas, N., Hadjigeorgiou, G. M., Papadimitriou, A., Dubois, B., Sarazin, M., Brandt, J., et al. (2004). Motor signs during the course of Alzheimer disease. *Neurology* 63, 975–982. doi:10.1212/01.wnl.0000138440.39918.0c

Schlink, B. R., Peterson, S. M., Hairston, W., König, P., Kerick, S. E., and Ferris, D. P. (2017). Independent component analysis and source localization on mobile eeg data can identify increased levels of acute stress. *Front. Hum. Neurosci.* 11, 310. doi:10.3389/fnhum.2017.00310

Seidl, U., Lueken, U., Thomann, P. A., Kruse, A., and Schröder, J. (2012). Facial expression in Alzheimer's disease. *Am. J. Alzheimers Dis. Other Demen.* 27, 100–106. doi:10.1177/1533317512440495

Sharma, K. (2019). Cholinesterase inhibitors as Alzheimer's therapeutics (Review). *Mol. Med. Rep.* 20, 1479–1487. doi:10.3892/mmr.2019.10374

Slegers, A., Filiou, R.-P., Montembeault, M., and Brambati, S. M. (2018). Connected speech features from picture description in Alzheimer's disease: a systematic review. *J Alzheimers Dis* 65, 519–542. doi:10.3233/jad-170881

Smith, M. (1995). Facial expression in mild dementia of the Alzheimer type. *Behav. Neurol.* 8, 149–156.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Front. Aging Neurosci.* 7, 195. doi:10.3389/fnagi.2015.00195

Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2020). Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *PLoS ONE* 15, e0236009. doi:10.1371/journal.pone.0236009

Tiele, A., Wicaksono, A., Daulton, E., Ifeachor, E., Eyre, V., Clarke, S., et al. (2020). Breath-based non-invasive diagnosis of Alzheimer's disease: a pilot study. *J. Breath Res.* 14, 026003. doi:10.1088/1752-7163/ab6016

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., et al. (2015). "Automatic detection of mild cognitive impairment from spontaneous speech using ASR," in Annual Conference of the international speech communication association.

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer. Res.* 15, 130–138. doi:10.2174/1567205014666171121114930

Villemagne, V. L., Doré, V., Burnham, S. C., Masters, C. L., and Rowe, C. C. (2018). Imaging tau and amyloid-β proteinopathies in Alzheimer disease and other conditions. *Nat. Rev. Neurol.* 14, 225–236. doi:10.1038/nrneurol.2018.9

Voleti, R., Liss, J. M., and Berisha, V. (2019). A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE J. Selected Top. Signal Process.* 14, 282–298. doi:10.1109/JSTSP.2019.2952087

Walker, T., Christensen, H., Mirheidari, B., Swainston, T., Rutten, C., Mayer, I., et al. (2020). Developing an intelligent virtual agent to stratify people with cognitive complaints: a comparison of human-patient and intelligent virtual agent-patient interaction. *Dementia* 19, 1173–1188. doi:10.1177/1471301218795238

Wang, S.-J., Liao, K.-K., Fuh, J.-L., Lin, K.-N., Wu, Z.-A., Liu, C.-Y., et al. (1994). Cardiovascular autonomic functions in Alzheimer's disease. *Age Ageing* 23, 400–404. doi:10.1093/ageing/23.5.400

Werner, P., Rosenblum, S., Bar-On, G., Heinik, J., and Korczyn, A. (2006). Handwriting process variables discriminating mild Alzheimer's disease and mild cognitive impairment. *J. Gerontol. Ser. B: Psychol. Sci. Soc. Sci.* 61, P228–P236. doi:10.1093/geronb/61.4.p228

Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., Prina, A. M., Winblad, B., et al. (2017). The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's Demen.* 13, 1–7. doi:10.1016/j.jalz.2016.07.150

World Health Organization and Alzheimer's Disease International (2012). *Dementia: a public health priority*. London: World Health Organization.

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. *Proc. Interspeech* 12, 2162–2166. doi:10.21437/interspeech.2020-2516

Zhang, Y., Wilcockson, T., Kim, K. I., Crawford, T., Gellersen, H., and Sawyer, P. (2016). Monitoring dementia with automatic eye movements analysis. *Intell. Dec. Technol.* 11, 299–309. doi:10.1007/978-3-319-39627-9_26

Zhou, Q., Srivastava, N., Goncalves, J., Newn, J., Dingler, T., and Velloso, E. (2019). "Cognitive aid: task assistance based on mental workload estimation," in ACM HCI Conference on human Factors in computing systems May 2019, Glasgow Scotland UK, (Association for Computing Machinery), 1–6. doi:10.1145/3290607.3313010

Zulli, R., Nicosia, F., Borroni, B., Agosti, C., Prometti, P., Donati, P., et al. (2005). QT dispersion and heart rate variability abnormalities in Alzheimer's disease and in mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 2135–2139. doi:10.1111/j.1532-5415.2005.00508.x

Conflict of Interest: The 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.

frontiers
in Aging Neuroscience

ORIGINAL RESEARCH
published: 27 April 2021
doi: 10.3389/fnagi.2021.637404

Cognitive and Structural Correlates of Conversational Speech Timing in Mild Cognitive Impairment and Mild-to-Moderate Alzheimer's Disease: Relevance for Early Detection Approaches

Céline De Looze[1], Amir Dehsarvi[1], Lisa Crosby[2], Aisling Vourdanou[1], Robert F. Coen[2], Brian A. Lawlor[2,3] and Richard B. Reilly[1,3]*

[1] Trinity Centre for Biomedical Engineering, School of Engineering, Trinity College Dublin, Dublin, Ireland, [2] Mercer's Institute for Successful Ageing, St James's Hospital, Dublin, Ireland, [3] Institute of Neuroscience, School of Medicine, Trinity College Dublin, Dublin, Ireland

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Sofia De La Fuente Garcia,
University of Edinburgh,
United Kingdom
David Glenn Clark,
Indiana University, Purdue University,
Indianapolis, United States

***Correspondence:**
Céline De Looze
deloozec@tcd.ie

Received: 03 December 2020
Accepted: 31 March 2021
Published: 27 April 2021

Citation:
De Looze C, Dehsarvi A, Crosby L, Vourdanou A, Coen RF, Lawlor BA and Reilly RB (2021) Cognitive and Structural Correlates of Conversational Speech Timing in Mild Cognitive Impairment and Mild-to-Moderate Alzheimer's Disease: Relevance for Early Detection Approaches. Front. Aging Neurosci. 13:637404. doi: 10.3389/fnagi.2021.637404

Background: Increasing efforts have focused on the establishment of novel biomarkers for the early detection of Alzheimer's disease (AD) and prediction of Mild Cognitive Impairment (MCI)-to-AD conversion. Behavioral changes over the course of healthy ageing, at disease onset and during disease progression, have been recently put forward as promising markers for the detection of MCI and AD. The present study examines whether the temporal characteristics of speech in a collaborative referencing task are associated with cognitive function and the volumes of brain regions involved in speech production and known to be reduced in MCI and AD pathology. We then explore the discriminative ability of the temporal speech measures for the classification of MCI and AD.

Method: Individuals with MCI, mild-to-moderate AD and healthy controls (HCs) underwent a structural MRI scan and a battery of neuropsychological tests. They also engaged in a collaborative referencing task with a caregiver. The associations between the conversational speech timing features, cognitive function (domain-specific) and regional brain volumes were examined by means of linear mixed-effect modeling. Genetic programming was used to explore the discriminative ability of the conversational speech features.

Results: MCI and mild-to-moderate AD are characterized by a general slowness of speech, attributed to slower speech rate and slower turn-taking in conversational

Abbreviations: AD, Alzheimer's disease; Cereb, cerebellum; CGP, cartesian genetic programming; EAs, evolutionary algorithms; FFG, fusiform gyrus; HC, healthy controls; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; L, left; MFG, middle frontal gyrus; MTG, middle temporal gyrus; MCI, mild cognitive impairment; Prec, precuneus; R, right; ROI, region-of-interest; STG, superior temporal gyrus.

settings. The speech characteristics appear to be reflective of episodic, lexico-semantic, executive functioning and visuospatial deficits and underlying volume reductions in frontal, temporal and cerebellar areas.

Conclusion: The implementation of conversational speech timing-based technologies in clinical and community settings may provide additional markers for the early detection of cognitive deficits and structural changes associated with MCI and AD.

Keywords: speech timing, conversation, cognitive function, brain volumes, Alzheimer

INTRODUCTION

Rationale and Research Goals

Progressive loss of cognitive function and progressive cerebral atrophy are characteristic features of Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD; Dubois et al., 2007; McKhann et al., 2011; Kirova et al., 2015; König et al., 2015; Szatloczki et al., 2015). Early and cost-effective diagnosis is crucial for the development and establishment of early interventions and to make effective treatment decisions.

Increasing efforts have focused on the establishment of novel biomarkers for the early detection of AD and prediction of MCI-to-AD conversion, including clinical, brain, genetic, and neuropsychological data. Behavioral changes over the course of healthy ageing, at disease onset and during disease progression, have been recently put forward as promising markers for the detection of MCI and AD. Repeated behavioral measures taken from everyday situations (e.g., walking speed) and/or extracted from tests that can be easily implemented outside clinical settings may offer the opportunity to increase timely detection and represent additional sources to the standard brain imaging and clinical neuropsychological assessments (e.g., MMSE—Mini Mental State Examination).

Speech-based approaches have proved to perform well in the discrimination of MCI and AD (König et al., 2015; López-de-Ipiña et al., 2015; Weiner et al., 2016; De Looze et al., 2018; Mirheidari et al., 2020). Speech and language impairments are indeed salient characteristics of MCI and early AD (Ripich et al., 1991; Caramelli et al., 1998; Chapman et al., 2002; Carlomagno et al., 2005; Taler and Phillips, 2008; Laws et al., 2010; Gayraud et al., 2011; Ahmed et al., 2013). However, the cognitive and structural underpinnings of these speech-based measures in classification approaches have not been systematically investigated and are not fully established. Understanding these underpinnings could add significant clinical value and further support the potential use and implementation of speech-based technologies in and outside clinical settings for the monitoring of cognitive trajectories.

One candidate tool is the analysis of spontaneous speech in conversational interactions. Engaging in a conversation is a complex skill which requires the integration of multiple independent cognitive subsystems, themselves supported by extensive networks of several brain regions. If one of these subsystems or networks is impaired, conversational speech difficulties may arise. Conversational speech characteristics may therefore be sensitive markers of underlying cognitive and structural impairments. The present study examines whether the temporal organization of speech in a collaborative referencing task is associated with cognitive function and the volumes of brain regions involved in speech production and known to be reduced in MCI and AD pathology. We then explore the discriminative ability of the temporal speech measures for the classification of MCI and AD.

Speech-Based Approaches for the Early Detection of MCI and AD

The potential use of speech-based approaches for the early detection of MCI and AD represents an important line of research in AD speech pathology. Deficits in the lexical, semantic, executive, discourse and pragmatic domains of language are commonly observed in MCI and early AD (Ripich et al., 1991; Caramelli et al., 1998; Chapman et al., 2002; Carlomagno et al., 2005; Feyereisen et al., 2007; Taler and Phillips, 2008; Laws et al., 2010; Gayraud et al., 2011; Ahmed et al., 2013; Drummond et al., 2015; Mueller et al., 2018). Symptoms include word-finding difficulties, decreased semantic and phonemic fluency, lexical richness, syntactic complexity and topic coherence. They often occur before clinical diagnosis and progress over the course of the disease (Ahmed et al., 2017). The articulatory aspects of language production are generally preserved until the late stages of the disease (Croot et al., 2000). Several speech and language tests and measures have been employed for the classification of MCI and AD, with accuracy rates spanning from 0.71 to 0.80 for the discrimination of MCI vs. healthy controls (HCs) and 0.80–0.98 for the AD vs. HC contrast (Roark et al., 2011; Jarrold et al., 2014; Meilán et al., 2014; König et al., 2015; López-de-Ipiña et al., 2015; Dodge et al., 2015; Asgari et al., 2017; Tóth et al., 2018; Gosztolya et al., 2019; O'Malley et al., 2020).

Speech Timing in Conversational Speech: Cognitive Underpinnings

Engaging in a conversation is a complex skill which requires the integration and coordination of multiple independent cognitive processes as speakers perform a number of tasks simultaneously. They must comprehend their interlocutor's utterances while, at the same time, prepare their response, keep track of the conversation topic, of the interlocutor's intent, and anticipate turns ending (Sacks et al., 1978; Riest et al., 2015). Smoothed exchanges of turns rely on the good functioning of a number of different cognitive abilities, including lexical and semantic retrieval, episodic memory, sustained attention, working memory, executive function, and language comprehension (Mueller et al., 2018).

Deficits in every one of these domains and conversational speech and language difficulties have been documented in adults with MCI and AD (Carlomagno et al., 2005; Feyereisen et al., 2007; Taler and Phillips, 2008; Rousseaux et al., 2010; Forbes-McKay et al., 2013; Drummond et al., 2015; Fraser et al., 2016). Difficulties in understanding words and sentences and producing words have been attributed to impairments in lexical and semantic retrieval (Murdoch et al., 1987; Forbes-McKay et al., 2013). Difficulties in discourse organization and turn-taking management are thought to stem from deficits in executive functioning (Rousseaux et al., 2010; Ash et al., 2012).

The temporal aspects of conversational speech within the frame of turn-taking organization may be a particularly sensitive marker of an individual's cognitive capacity. Analyses of connected speech revealed that AD speech is characterized by slower speech rate (global speed of speech including pauses), a higher number of silent pauses, longer pauses and shorter interpausal units (or chunks of speech bounded by silent pauses; Weiner et al., 2008; Davis and Maclagan, 2009; Rousseaux et al., 2010; Hoffmann et al., 2010; Gayraud et al., 2011; Pistono et al., 2016; De Looze et al., 2018). Slower speech rate, a higher number of silent pauses, a reduction in phrase length and an increase in speech turns frequency were also observed in MCI and AD conversational speech (Carlomagno et al., 2005; Hoffmann et al., 2010; Sajjadi et al., 2012).

Slower speech rate, larger pause frequency, and longer pause duration have been mainly attributed to lexico-semantic deficits in MCI and AD (Goldman Eisler, 1968; Hoffmann et al., 2010; Forbes-McKay et al., 2013; Pistono et al., 2016). Other studies have also pointed towards further deficits in working memory, attention, and executive function (Ash et al., 2012; Pistono et al., 2016; De Looze et al., 2018). Longer pauses between clauses have been associated with speech planning difficulties (Matsumoto et al., 2013). In addition, the manner in which readers chunks their speech stream into units of different sizes was shown to be dependent on their working memory (WM) capacity. In healthy older adults, readers with low WM capacity were more likely to chunk their speech into smaller units than those with high WM, indicating a narrower scope of planning (Ferreira and Swets, 2002; Swets et al., 2014). In a previous study, we found that, in overt sentence reading, a higher number of pauses, shorter interpausal units and slower speech rate were associated with reduced language and working memory/attention scores and that these temporal speech characteristics were reflective of difficulties in planning longer and more syntactically complex utterances in healthy older adults and individuals with MCI and AD (De Looze et al., 2018).

Together these separate findings suggest that the temporal organisation of speech in MCI and AD may be indicative of a number of underlying cognitive deficits, e.g., deficits in episodic memory, lexical retrieval, executive functions, working memory and attention. However, these associations are not well established within the frame of conversational interactions.

Speech Timing in Conversational Speech: Structural Correlates

During conversational interactions, several brain regions are recruited and formed into extensive networks to support visual, phonological, lexical, semantic, syntactic, pragmatic, discourse, and attentional processes.

Besides a limited number of studies describing the neural correlates of conversational speech production, a number of regions are thought to be involved in these cognitive processes. A widespread distribution of language areas in the temporal, parietal, and frontal lobes have been associated with lexical-semantic memory and retrieval (Binder et al., 2009). Naming performance has been associated with the left anterior temporal lobe, including the left temporal pole, the left inferior temporal gyrus (ITG), the left middle temporal gyrus (MTG), the left superior temporal gyrus (STG), and the left fusiform gyrus (L FFG; Kircher et al., 2004; Brambati et al., 2006; Binder et al., 2009; Baldo et al., 2013; Pravatà et al., 2016; Leyton et al., 2019). Involvement of the left inferior parietal gyrus (Kircher et al., 2004; Baldo and Dronkers, 2006) and the left inferior frontal gyrus (IFG; Binder et al., 2009; Hurley et al., 2015) has also been reported. The temporal regions are thought to be related to the activation and storage of lexical representations while the frontal areas have been linked specifically to the retrieval aspect of lexico-semantic processing (Hagoort, 2005; Binder et al., 2009).

Language areas associated with speech planning, executive functions and, more specifically, the monitoring of turn-taking organization, include the motor cortex, the middle and inferior frontal gyri, the inferior parietal lobule (IPL), and the STG (Hagoort, 2005; Matsumoto et al., 2013; Magyari et al., 2014; Foti and Roberts, 2016; Nissim et al., 2017). The left IFG is thought to support the parsing and planning of sentence and discourse-level linguistic information (Matsumoto et al., 2013; Magyari et al., 2014). The midfrontal areas have been related to verbal action planning and attentional control (Hagoort, 2005) and the IPL has been linked to verbal working memory capacity (Deschamps et al., 2014). These regions together are thought to play a central role in sentence and discourse level comprehension processes and control, particularly in turn-ending anticipation (Magyari et al., 2014). Other regions reported to be associated with working memory and executive function in speech processing and production include the precuneus/posterior cingulate cortex and the cerebellum (Cereb; Xu et al., 2005; Hampson et al., 2006; Newman et al., 2013; Bourguignon, 2014; Christodoulou et al., 2014; Hirshorn et al., 2014; Helder et al., 2017). Increased activation of these two regions together with the IFG, MTG, and IPL have been related to working memory capacity in sentence reading and comprehension, potentially reflecting the additional working memory demand that emerges at the sentential/discourse level (Xu et al., 2005; Prat et al., 2007; Newman et al., 2013; Helder et al., 2017; De Looze et al., 2018).

Pauses within clauses, reflective of lexico-semantic processes, have been associated with activation in the superior and middle temporal gyri bilaterally (Kircher et al., 2004). Between-clause pauses, reflective of speech planning and monitoring, have been related to the left STG, the left insula, and the right IFG (Kircher et al., 2004; Matsumoto et al., 2013). Inter-speaker gaps (i.e., the silence between two speakers' turns), underlying the anticipation of a speaker's response, have been associated with the posterior temporal gyrus, the supramarginal gyrus, the premotor cortex and middle prefrontal cortex (Bögels et al., 2015; Foti and Roberts, 2016). Speech rate, reflective of speech motor control and planning, has been related to the STG bilaterally, the left MTG, the right ITG, the right fusiform gyrus (R FFG), the left and right IPL, and the precuneus (Ash et al., 2012).

Widespread changes in the structure, function, and organization of a multitude of brain regions have been reported in MCI and AD. Beyond a typical atrophy of the medial temporal lobe (Lehéricy et al., 1994; Chan et al., 2001; Dickerson et al., 2001; Killiany et al., 2002), volume reductions in the fusiform gyrus (FFG), posterior cingulate/precuneus, superior temporal, inferior parietal, and orbito-frontal cortices were also observed in MCI and AD (Tondelli et al., 2012; Wang et al., 2015; Dicks et al., 2018; Verfaillie et al., 2018). Given the overlap of regions engaged in speech processing and production and reduced in MCI and AD pathology, it may be hypothesized that conversational speech timing characteristics may be reflective of underlying regional volume reductions. Evidence in the context of conversational interactions is however limited.

Objectives and Hypotheses

In this study, we first examine whether the temporal organization of conversational speech in a collaborative referencing task is associated with cognitive function in individuals with MCI and AD. In a second analysis, we investigate whether conversational speech timing is reflective of the underlying volume of brain regions involved in speech production and known to be reduced in MCI and AD pathology. We consider an extensive ensemble of conversational speech timing measures, cognitive domains and brain regions. We expected shorter interpausal units, shorter turns, longer pauses, longer gaps, shorter transition overlaps, a higher number of pauses and gaps and slower speech rate to be associated with lower cognitive function and reduced regional brain volumes. These analyses aim to establish which conversational speech measures reflect underlying cognitive deficits and regional brain volume reductions in order to estimate their clinical relevance for the implementation of speech-based technologies for the monitoring of speech changes in healthy ageing, MCI and AD. Finally, we explore the discriminative ability of these temporal speech measures for the classification of MCI and AD using Cartesian genetic programming (CGP). Although our analyses are exploratory due to the sample size under investigation, to our knowledge, this is the first attempt to examine the discriminative ability of conversational speech measures while also investigating their cognitive and structural underpinnings using the same cohort.

MATERIALS AND METHODS

Ethics Statement

Ethical approval for the study was obtained from the St. James's Hospital Ethics and Medical Research Committee. Signed informed consent was obtained from all respondents prior to participation.

Participants

Twenty older adults with MCI and 20 older adults with mild-to-moderate AD were recruited from the Memory Clinic of the Mercers Institute for Successful Ageing (MISA) in St. James's Hospital, Dublin, Ireland. Forty healthy volunteers (HC) were recruited from the Memory Research Unit in Trinity College Dublin. Participants included in this study were over 50 years of age, fluent in English and literate, to ensure that they could complete all assessments and tasks ($N = 80$). MCI and mild-to-moderate AD diagnoses were based on NIA-AA criteria (Albert et al., 2011; Sperling et al., 2011). Exclusion criteria for healthy participants included history of neurological disorders and/or history of major psychiatric disorders or depression. Thirteen participants with MCI, 13 with AD and 16 HC, without prior MRI contraindications, e.g., pacemakers, cerebral aneurysm clips or other, were randomly selected to undergo brain MRI ($N = 42$). Three participants with MCI, three with AD and four HC were excluded from analysis due to incomplete MRI scans, technical issues with the MRI data (e.g., motion artefact, volume segmentation errors), technical issues with the speech data (e.g., recording issues) and/or abnormal scans or cognitive scores for the HC. Participants with AD, MCI and HC who had reliable cognitive, speech and MRI volumetric measures were included for analysis, resulting in a final sample of 32 individuals.

Neuropsychological Tests

All participants underwent two neuropsychological tests, which were administered and assessed by an experienced nurse. The RBANS (Repeatable Battery for the Assessment of Neuropsychological Status; Randolph et al., 1998) includes five cognitive domains: verbal memory (immediate and delayed recall), visuospatial/constructional abilities (figure copy and orientation), attention (symbol and digit coding), working memory (forward and backward digit span) and language (naming and semantic fluency). The MoCA (Montreal Cognitive Assessment; Nasreddine et al., 2005) is composed of 14 tests subsumed under six different cognitive domains which include visuoconstructional/executive function skills (figure copy, clock drawing, and trail test), verbal memory (delayed recall), attention/working memory (sustained attention, serial 7s, forward and backward digit span), language (naming, sentence repetition and phonemic fluency), conceptual thinking (verbal abstraction), and orientation (time and place). Composite scores of five different cognitive domains were computed from age, gender, and education-corrected

RBANS and MOCA raw scores, by averaging them for each specific cognitive domain as previously described (De Looze et al., 2018): *Memory* was generated from RBANS and MOCA immediate and delayed recall scores; *Language* from RBANS/MOCA Language and MOCA naming scores; *Working Memory/Attention* from RBANS/MOCA working memory and attention scores; *Visuoconstructional/Executive function* from the RBANS Visuospatial/Constructional scores and MOCA Visuoconstructional/Executive function scores; *Orientation* from MOCA orientation index. All cognitive tests took place in the Memory Clinic of the MISA in St. James's Hospital. The tests took on an average 35–55 min to complete.

Collaborative Referential Task

Participants were asked to engage in a collaborative referential task (Feyereisen et al., 2007; Duff et al., 2013) with a communication partner. The communication partner was the caregiver of the participants with MCI/AD. Each caregiver engaged twice in the referential task, once with the participants with MCI/AD and once with a matched HC. Each pair of individuals engaged in three trials.

In the first trial (*Describe*-Trial), individuals with AD, MCI or the HCs were the directors and the caregivers were the matchers. The directors were given a board with 10 numbered spaces and a set of 10 cards displaying Chinese tangrams arranged on the board in a unique sequence. The matchers were given an identical board with 10 numbered spaces and an identical set of cards which were randomly displayed around the board. The tangrams were black and white geometric shapes which could resemble human beings, animals, or objects but which had no established names. The directors were asked to describe the shapes and tell the matcher where to place them on their board so that, at the end of the trial, the director's and the matcher's boards looked alike. In the second trial (*Match*-Trial), the roles were inversed. In the third trial (*Describe and Match*-Trial 3), the pair had to discuss together the identical shapes that they were given and agree on where to place them on their respective boards, so that at the end of the trial their boards matched.

The sets of Chinese tangrams were different for each trial, but the same sets were used across pairs of individuals. During the task, the pairs were seating at a table facing each other. A partial barrier or stand up obscured the view of the other's board, facial expressions and gestures to rely on speech communication only. The task was presented as a game and the participants were told to have fun. The experimenter was siting aside in the room working on a computer while the pairs played the game. Feedback about the total number of correct card placements was provided after each trial.

All sessions took place in a clinical room located in the Memory Clinic and were audio-recorded. H4n Zoom recorders were used for the recordings. The audio signal was recorded at 44 kHz/16 Bit resolution. The collaborative referential task (three trials) took on average 12 min.

Speech Annotation and Measure Extraction

The conversational speech data was annotated using the Praat software (Boersma and Weenink, 2016). Speech units and silences were first automatically determined, using a binary voice activity detection (VAD) algorithm proposed in Sohn et al. (1999). Turns, interpausal units, pauses, gaps, and transition overlaps were then automatically derived from the binary VAD using Praat scripts. Interpausal units are speech units separated by a pause. A turn is defined herein as a unit of speech composed of one or several consecutive interpausal units produced by the same speaker. Pauses are silences within a speaker's turn. The pause threshold used in the automatic procedure was set at 100 ms to ensure its distinction with silent plosives (Sanderman and Collier, 1995). Gaps are silences at turn boundaries, that is when there is a change in speaker. Transition overlaps denote chunks of speech when two speakers speak simultaneously at turn boundaries. Syllables were automatically aligned to the signal using a modified version of de Jong and Wempe's (2009) Praat script. The acoustic annotation was manually checked by a speech expert and corrected where needed.

Speech timing measures were automatically extracted using Praat scripts and included *each* participant (AD/MCI/HC)'s total number of pauses; gap/transition overlap ratio; duration of pauses, gaps, transition overlaps, interpausal units, and turns; and speech rate (number of syllables per second including pauses). The number of pauses were normalized to the speaker's turn. The gap/transition overlap ratio was calculated as the number of gaps divided by the number of overlaps. The higher the ratio, the higher the tendency for an individual to use a gap (rather than a transition overlap) when taking a turn.

MRI Protocol and T1w Acquisition

Participants were scanned at the National Centre for Advanced Medical Imaging (CAMI), St. James' Hospital, Dublin, using a 3T Philip's Achieva system and 32-channel head coil. A 3D Magnetization Prepared Rapid Gradient Echo (MP-RAGE) sequence was used to acquire various scans in addition to a T1-weighted MR image. Scans included the subsequent parameters: FOV (mm): 240 × 218 × 162; 0.9 mm isotropic resolution; SENSE factor: 2; TR: 2 ms; TE: 2.8 ms; flip angle: 8°. The MRI data was obtained within one to 3 weeks after the cognitive and speech assessments.

MRI Data Inspection

FreeSurfer software version 6.0 (Dale et al., 1999) was used to analyze the T1w images with the associated cross-sectional pipeline to derive Regions of Interest (ROIs) in each subject's native space, using the Destrieux atlas (Destrieux et al., 2010). The technical details of FreeSurfer procedures have been described elsewhere (Dale et al., 1999; Fischl et al., 2002; Han et al., 2006; Jovicich et al., 2006). All unprocessed input volumes were inspected for evidence of motion artefact. Surface segmentation failures were identified using Freeview.

Feature Extraction

We selected nine regions of interest (ROIs) which were found to be involved in speech production (Xu et al., 2005; Hampson et al., 2006; Newman et al., 2013; Bourguignon, 2014; Christodoulou et al., 2014; Hirshorn et al., 2014; Helder et al., 2017) and reduced in MCI and AD pathology (Tondelli et al., 2012; Wang et al., 2015; Dicks et al., 2018; Verfaillie et al., 2018): the IFG (the sum of the pars opercularis, pars triangularis and pars orbitalis), the Middle Frontal Gyrus (MFG), the Precuneus (Prec), the IPL (the sum of the Angular Gyrus and the Supra Marginal Gyrus), the ITG, the MTG, the planum temporale in the STG, the FFG, and the Cereb. The volumes of these regions were extracted from FreeSurfer cortical segmentation statistical output. Measures were obtained separately for each hemisphere, which resulted in a total of 18 ROIs. Total Gray Matter volume was extracted to assess Group differences. Estimated Total Intracranial Volume served to control for individual differences in head size in regression analyses.

Statistical Analyses

All statistical analyses were performed using R software version 3.6.0 (R Foundation for Statistical Computing, Vienna, Austria; R Core Team, 2015).

Data Descriptives

The observed sample was first characterized per Group (HC, MCI, and AD). Continuous variables were described as the mean with standard deviation; categorical variables were given as percentage. Ordinary least square and generalized models, when appropriate, were used for comparison of demographics, neuropsychological scores and speech task competence. HC was set as the reference level.

Speech Timing by Group

The effects of Group on the temporal characteristics of speech were assessed through a mixed model approach (Bates et al., 2014). Linear mixed effects models are an extension of simpler linear models. They include both fixed and random effects as predictor variables. They are robust for the analysis of repeated measures designs and can account for both within- and between-subject factors (Littell et al., 1996).

Ten temporal characteristics of speech (i.e., the number of pauses, gap/transition overlap ratio, the duration of pauses, gaps, transition overlaps, interpausal units, and turns, and speech rate) were entered as *dependent variables* (repeated measures) in separate linear mixed-effect models. The dependent variables were log-transformed when appropriate. For all models, *fixed effects* were the Group (AD, MCI, and HC) and Trial (*Describe*-Trial, *Match*-Trial, and *Describe and Match*-Trial), with an interaction term. HC Group and *Describe and Match* -Trial were set as the reference levels. Trial was included as a fixed effect to reflect the participant's role, hence the different cognitive load and speech task involved in each trial. Speakers constituted the *random intercepts*. All our models were adjusted for age, sex, and education.

The significance of interaction terms was assessed through likelihood ratio tests comparing additive models with models with an interaction term. The significance level was set at $\alpha = 0.006$ to correct for multiple comparison ($\alpha = 0.05/8$ models). Following standard procedures for mixed models (Nakagawa and Schielzeth, 2013), both marginal (R^2m, describing the proportion of variance explained by the fixed factors alone) and conditional (R^2c, describing the proportion of variance explained by both the fixed and random factors) R^2 were computed to assess effect size.

Association Between Speech Timing and Composite Scores

The association between the temporal characteristics of speech and the composite scores were assessed using linear mixed effects models. As per above, the number of pauses, gaps and transition overlaps (normalized), the gap/transition overlap ratio, the duration of pauses, gaps, transition overlaps, interpausal units, and turns and speech rate were entered as *dependent* variables (repeated measures) in separate models. The dependent variables were log-transformed when appropriate.

In each model, the five composite scores Working Memory/Attention, Language, Memory, Visuoconstructional/Executive Function and Orientation (continuous variables) and Trial (three levels: *Describe*-Trial, *Match*-Trial and *Describe and Match*-Trial) were entered as *fixed effects*, with an interaction term. *Describe and Match*-Trial was set as the reference level. A stepwise procedure (backward and forward) was employed to assess the significance of the predictors. The composite scores were centered around their mean to reduce multicollinearity.

In all models, speakers constituted the *random intercepts*. All our models were adjusted for age, sex, and education. The significance level in the full models was set at $\alpha = 0.006$ to correct for multiple comparison ($\alpha = 0.05/8$ models). Marginal (R^2m) and conditional (R^2c) R^2 were used to estimate effect size.

Association Between Speech Time and Regional Volumes

A biologically informed ROI-based approach was chosen to explore the association between the temporal characteristics of speech and regional volumes through linear mixed-effect modelling (Bates et al., 2014). The number of pauses, gaps and transition overlaps (normalized), the gap/transition overlap ratio, the duration of pauses, gaps, transition overlaps, interpausal units, and turns and speech rate were entered as *dependent* variables (repeated measures) in separate models. The dependent variables were log-transformed when appropriate.

In each model, the ROIs (z-score transformed) and Trial were entered as *fixed effects*, with an interaction term. ROIs of the left and the right hemispheres were run separately. *Describe and Match*-Trial was set as the reference level. A stepwise procedure (backward and forward) was employed to assess the significance of the predictors. Speakers constituted the *random intercepts*. All our models were adjusted for age, sex, and education. The significance level in the full models was set at $\alpha = 0.003$ to correct for multiple comparison ($\alpha = 0.05/16$ models). Marginal (R^2m) and conditional (R^2c) R^2 were used to estimate effect size.

Classification of MCI and AD Based on Speech Features

Finally, we explored the discriminative ability of temporal speech characteristics and the use of CGP, a subtype of Evolutionary Algorithms (EAs), for the classification of MCI and AD.

Rationale for Using Cartesian Genetic Programming. EAs are learning algorithms derived from Darwinian evolutionary theory. CGP is a subtype of EAs, which generates directed acyclic computational configurations of nodes. Like other types of EAs, it uses trees as its solution representation

(Miller, 2020). CGP can evolve symbolic expressions, Boolean logic circuits, and artificial neural networks. The algorithms generate a population of classifiers through a repeated process of variation and selection. Selection is based on improving fitness criteria when categorizing the participant groups from each other. EAs are stochastic (i.e., different solutions are found each time the algorithms are executed), hence, in order to address this, the best performing classifier was selected from numerous repeated runs of the algorithms. Unlike other standard mathematical approaches and most machine learning algorithms, CGP, like other EAs, makes very few assumptions about the function that generated the data, which allows a wide exploration of the space possible solutions to the problem. The general scheme of an EA is presented in **Figure 1**.

Computational methods, such as EAs, have been recently used for the measurement and analysis of clinical data (e.g., patient movements data and neuroimaging, among others; Dehsarvi and Smith, 2018). A core advantage when applying EAs with an expressive dynamical representation is that multiple classifiers can be examined. In addition, EAs offer a *white-box solution* for the classification, which is not the case for most (*black-box*) machine learning algorithms. With *white-box* models, the classification process is transparent; it is possible to retrieve how predictions were produced and which variables influenced the population and selection of classifiers. Upon the completion of the classification process, EAs allows for looking into the classification graphs generated by the algorithm and, for instance, exploring how specific features have been chosen to evolve the models. Finally, EAs have proved to perform well with relatively small datasets (Picardi et al., 2017; Dehsarvi and Smith, 2018; Muhamed et al., 2018). To our knowledge, our study is the first to investigate whether the use of EAs and conversational speech may enhance the classification of MCI and/or AD.

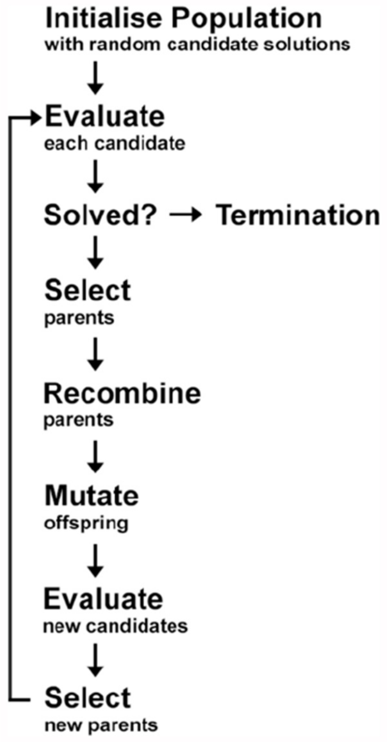


FIGURE 1 | General diagram of the classification process in Evolutionary Algorithms (EAs). In order to find the optimal model (or candidate), a set of working models are randomly generated (Step 1: initialize population). The models (or candidates) are then evaluated to assess their accuracy rate (Step 2: evaluate). In order to achieve the maximum accuracy rate, certain models (or candidates) are selected for use in the subsequent generation of models (Step 3: select) *via* recombination (also known as sexual reproduction or crossover) and/or mutation. Recombination is an operator that is applied to two or more selected models (the so-called parents or genotype or chromosomes), by mixing their genetic material (genes), to create one or more new models (the children or new chromosomes or offspring). Mutation is applied to one model (asexual reproduction) or two models (sexual reproduction) and results in one new model. This procedure is repeated for many iterations and the resulting model is evaluated each time (Step 4: evaluate) or until the desired accuracy rate is achieved at which stage a final optimal model is selected and the process is terminated (Step 5: termination). Adapted from Figure 4.5 of Dehsarvi (2018).

Classification. Classification analyses were performed using a novel open source cross platform CGP library (version 2.4; Turner and Miller, 2015). The number of pauses, gaps and transition overlaps, the gap/transition overlap ratio, the duration of pauses, gaps, transition overlaps, interpausal units and turns, and speech rate were used as input features. Per-speaker means and standard deviations of each normalized feature were computed for the three trials separately. Two-class (binary) classification was performed for the AD-HC and MCI-HC contrasts as well as multi-class classification of the three groups. To have equal class representation, the data from each class was randomly divided into subsets of 60% (training), 20% (validation), and 20% (test). The geometry of the programs in the population (referred to as chromosomes) has fifty nodes with a function set of four mathematical operations (+, −, ×, /), multiple inputs (according to the dataset), and one (either class 1 or class 0 for each binary combination of speaker groups) or multiple (one combination per speaker group) outputs. At each generation of classification, the fittest chromosome is selected, and the next generation is formed with its mutated versions (mutation rate = 0.1). Evolution stops when 15,000 iterations are reached. To obtain statistical significance, we completed the analysis for 10 runs for each combination of inputs and the result was calculated as the average of the accuracy rates over the runs. The results (the winning chromosome—an example is provided in **Figure 2**, the networks, and the accuracy values) were stored for each run individually.

Five-fold Cross-validation. A 5-fold cross-validation was then performed in 10 runs for each combination of inputs to evaluate accuracy and obtain statistical significance. The accuracy was averaged over the runs. An advantage of cross-validation is the production of independent test sets that increases reliability. With *5*-fold cross-validation, one (of *5*) subset is the test set, one subset is the validation set, and the other three subsets are training sets. These sets are alternated, so every set is used once for testing the data.

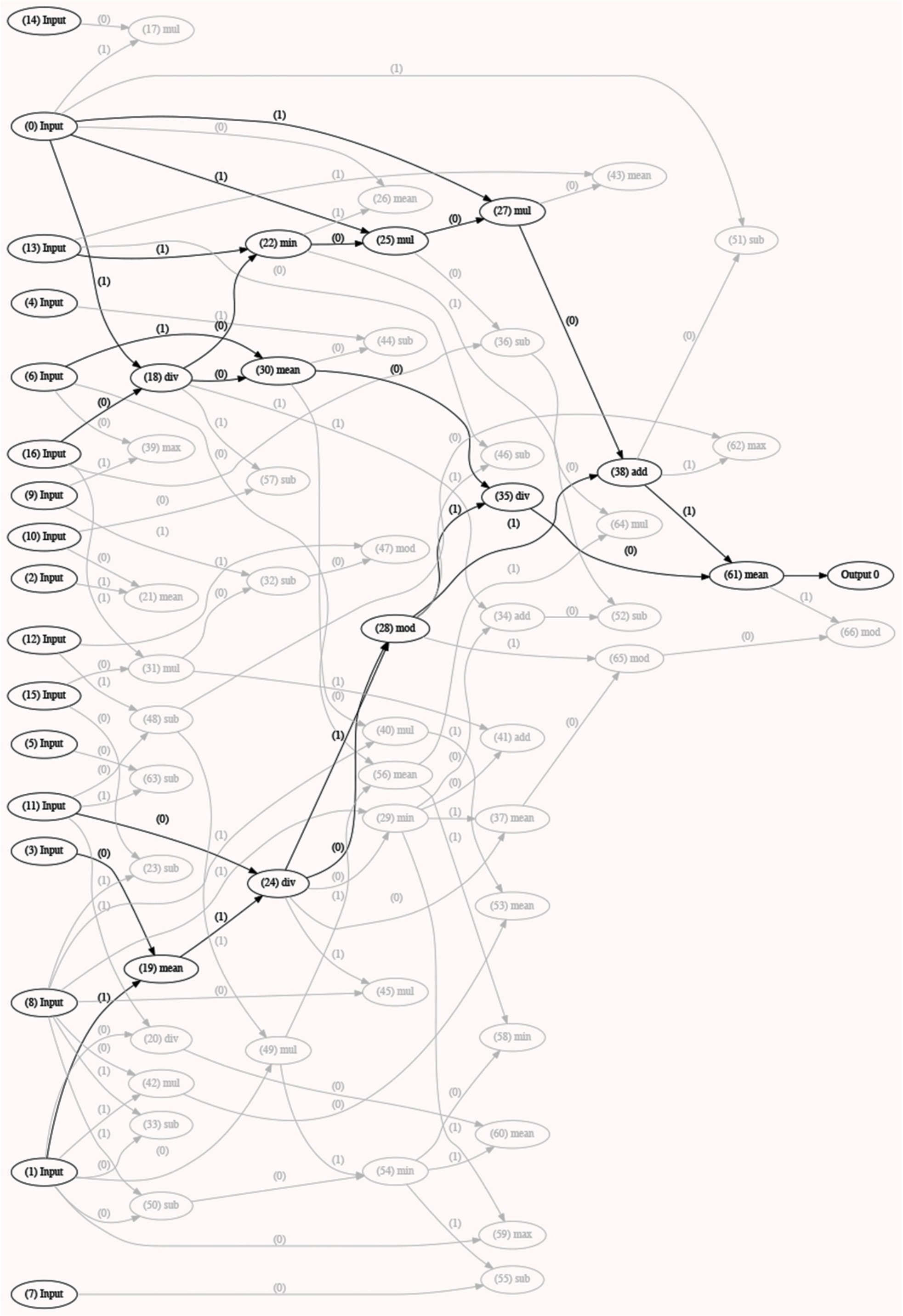


FIGURE 2 | Example of a generation of classification with the optimal model (best fitted chromosome) selected (in black). This model has used a certain number/set of inputs or speech features (inputs 0, 13, 6, 16, 11, 3, and 1) and a combination of different functions to form the best model (or fittest chromosome). Other models with lower accuracy rates are depicted in light gray in the figure. The selected model (or chromosome) is the fittest one of a certain run and is stored as an output, along with all the other runs, upon completion of 5-fold cross-validation.

One cycle of the 5-fold cross-validation does not generate enough classification accuracies to enable comparison, hence, in 5-fold cross-validation, this is repeated 10 independent times and mean accuracy across all the trials is calculated (with the data samples being randomly allocated in different sets). The results (the winning chromosome, the networks, and the accuracy values) were stored for each run individually and the test results over all the iterations were averaged and reported (**Table 6**).

RESULTS

Sample Characteristics

Table 1 provides descriptive statistics per group and the results from the least-square and generalized regressions. There was no statistical difference in age, gender or education level between the HC and the MCI or AD groups. Individuals with MCI and AD had lower RBANS and MOCA global scores compared to the HC (reference level). The Memory, Language, Working Memory/Attention, Visuoconstructional/Executive Function and Orientation composite scores were significantly lower for the AD group. MCI participants had lower Memory, Working Memory/Attention and Visuoconstructional/Executive Function composite scores. The AD group also had reduced Total Gray Matter volume.

Speech Timing by Group

Table 2 provides the mean and standard deviation of the speech characteristics per Group × Trial. Significant results ($p < 0.006$, i.e., after Bonferroni correction) and tendencies or marginally significant results ($p < 0.01$, i.e., after Bonferroni correction) are reported herein. Significant coefficients, 95% confidence intervals and R^2 are given for the three trials and per Trial when the interaction Group × Trial was significant ($p < 0.006$) in **Table 3**.

The interaction Group × Trial was significant for speech rate ($\chi^2_{(14)} = 19.15$, $p < 0.006$), interpausal unit duration ($\chi^2_{(14)} = 21.38$, $p < 0.006$) and turn duration ($\chi^2_{(14)} = 17.69$, $p < 0.006$).

Individuals with AD had significant slower speech rate in the *Describe*-Trial compared to HC ($p < 0.006$). They also produced shorter interpausal units ($p = 0.003$) across the three trials. Their transition overlaps tended to be shorter in the *Describe*-Trial ($p = 0.008$). MCI participants tended to produce longer turns in the *Describe*-Trial ($p = 0.01$). The gap/transition ratio tended to be larger for the MCI groups ($p = 0.009$) compared to the HC across the three trials, i.e., individuals with MCI used more often a gap than a transition overlap when taking a turn. There was no significant (or marginally significant) difference in the number of pauses between the AC or MCI and HC groups.

Together, these results suggest that AD participants speak more slowly and take a longer time to respond when engaged in a collaborative referential task compared to HC. Our findings also suggest that MCI participants tend to produce longer turns. Their response times to take turns also tended to be longer than HC.

Speech Timing—Domain-Specific Cognitive Function Association

Significant results ($p < 0.006$, i.e., after Bonferroni correction) and tendencies or marginally significant results ($p < 0.01$, i.e., after Bonferroni correction) are reported herein. Significant coefficients, 95% confidence intervals and R^2 are given for the three trials and per Trial when the interaction Group × Trial was significant ($p < 0.006$) in **Table 4**.

Speech Rate

The interaction Group × Trial was significant for the Memory ($\chi^2_{(11)} = 18.53$; $p < 0.006$) and Visuoconstructional/Executive function ($\chi^2_{(11)} = 19.22$; $p < 0.006$) components. Slower speech rate was significantly associated with lower Memory scores in the *Describe*-Trial ($p = 0.001$) and with lower Visuoconstructional/Executive function scores in the *Match*-Trial ($p = 0.004$).

Turn Duration

The interaction Group × Trial was significant for the Working Memory/Attention ($\chi^2_{(11)} = 19.22$; $p < 0.006$), Memory ($\chi^2_{(11)} = 18.53$; $p < 0.006$) and Visuoconstructional/Executive function components ($\chi^2_{(11)} = 14.57$; $p < 0.006$). Shorter turns were associated with lower Working Memory/Attention scores across the three trials ($p < 0.006$), with weaker associations for the *Describe*-Trial ($p < 0.006$) and in the *Match*-Trial ($p < 0.006$). Positive associations were also found with Memory scores in the *Match*-Trial ($p < 0.006$) and with Visuoconstructional/Executive function scores in the *Describe*-Trial ($p < 0.006$) and the *Match*-Trial (*marginal*, $p = 0.01$).

Interpausal Unit Duration

The interaction Group × Trial was significant for Memory ($\chi^2_{(11)} = 11.13$; $p = 0.003$) and marginally significant for Orientation ($\chi^2_{(11)} = 7.97$; $p = 0.01$). Shorter interpausal units tended to be associated with lower Memory scores ($p = 0.008$) and with lower Orientation scores ($p = 0.004$) in the *Describe*-Trial. Interpausal units also tended to be shorter with lower Working Memory/Attention scores across the three trials ($p = 0.009$). See **Figure 3**.

Pause Duration

The interaction Group × Trial was marginally significant for the Orientation component ($\chi^2_{(11)} = 9.83$, $p = 0.007$). Longer pauses tended to be associated with lower Orientation scores in the *Describe*-Trial ($p = 0.01$) and with lower Memory scores across the three trials ($p = 0.004$).

Gap/Transition Overlap Ratio

The interaction Group × Trial was marginally significant for the Language component ($\chi^2_{(11)} = 8.15$; $p = 0.01$). A larger ratio (i.e., a higher occurrence of gaps as compared to transition overlaps) tended to be associated with lower Language scores in the *Match*-Trial ($p = 0.006$).

Gap duration, transition overlap duration and the number of pauses were not associated with any of the cognitive domains.

To summarize, slower speech rate and shorter turns were significantly associated with lower Memory and Visuoconstructional/Executive Function scores, with shorter turns being further associated with lower Working Memory/Attention scores. Marginal associations suggest similar trends. Shorter interpausal units tended to be associated with lower Memory and Working Memory/Attention scores as well as with lower Orientation scores. Lower Memory and Orientation scores tended to be associated with longer pauses.

The Structural Correlates of Speech Timing

Significant results ($p < 0.003$, i.e., after Bonferroni correction) and tendencies or marginally significant results ($p < 0.006$, i.e., after Bonferroni correction) are reported herein. Significant coefficients, 95% confidence intervals and R^2 are given for the three trials and per Trial when the interaction Group × Trial was significant ($p < 0.003$) in **Table 5**.

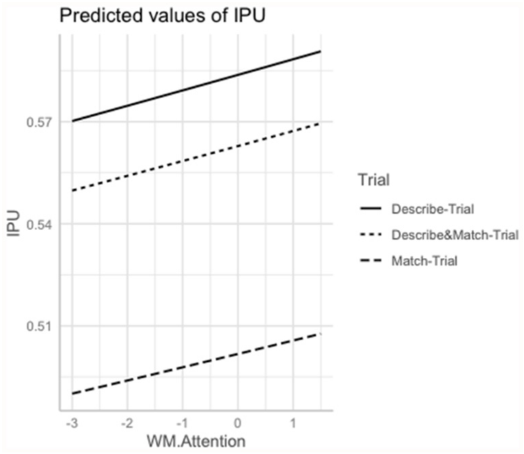


FIGURE 3 | Marginal estimates of Interpausal Unit (IPU) duration (log-transformed) as a function of Working Memory/Attention scores (log-transformed) in the *Describe*, *Match*, and *Describe and Match* trials. *Describe*-Trial: individuals with AD/MCI and HC are the directors, i.e., they describe the shapes and instruct where to place them; *Match*-Trial: individuals with AD/MCI and HC are the matchers, i.e., they place the pictures on their board following the caregiver's instructions; *Describe and Match*-Trial: both interlocutors describe the shapes and agree on where to place them.

Speech Rate

The interaction ROI*Trial was significant for the L MTG ($\chi^2_{(12)} = 13.54$; $p < 0.003$), R MTG ($\chi^2_{(12)} = 24.02$; $p < 0.003$) and R STG ($\chi^2_{(12)} = 12.84$; $p < 0.003$). In particular, slower speech rate was associated with smaller volume of L MTG and R MTG ($p < 0.003$) except in the *Match*-Trial; with smaller volume of R STG in the *Match*-Trial ($p < 0.003$).

Turn Duration

The interaction ROI*Trial was significant for the R MFG ($\chi^2_{(12)} = 10.36$; $p < 0.003$), and the R STG ($\chi^2_{(12)} = 9.57$; $p = 0.003$). Shorter turns were associated with smaller volumes of L MTG ($p < 0.003$) across the three trials. Shorter turns were also associated with smaller volume of R MFG ($p = 0.003$) in the *Match*-Trial and R STG ($p = 0.003$) in the *Describe*-Trial.

Interpausal Unit Duration

The interaction ROI*Trial was marginally significant for the L MTG ($\chi^2_{(12)} = 10.63$; $p = 0.004$), L ITG ($\chi^2_{(12)} = 10.22$; $p = 0.006$), and L Cereb ($\chi^2_{(12)} = 10.22$; $p = 0.007$). Shorter interpausal units were associated with smaller volume of L MTG in the *Describe*- and *Match* Trials ($p < 0.003$), with smaller volume of L ITG in the *Describe*-Trial ($p < 0.003$) and with smaller volume of L Cereb in the *Describe*-Trial ($p = 0.004$). The R IFG volume was also positively associated with interpausal unit duration, with weaker association in the *Describe*-Trial ($p = 0.003$) and *Match*-Trials ($p < 0.003$). See **Figure 4**.

Gap Duration

Gap duration was positively associated with the R ITG volume except in the *Describe* and *Match*-Trials ($p < 0.003$).

No association between pause duration, transition overlap duration, number of pauses, gap/transition overlap ratio and the ROIs volumes were found.

To summarize, slower speech rate, shorter turns and shorter interpausal units were significantly associated with smaller volumes of L MTG. Speech rate was also positively associated with the R MTG and R STG volumes; turn duration with the

TABLE 1 | Comparison of demographic and neuropsychological characteristics of participants with mild cognitive impairment (MCI), participants with mild-to-moderate Alzheimer's disease (AD) and healthy controls (HCs).

	AD (*N* = 10)	MCI (*N* = 10)	HC (*N* = 12)	AD vs. HC (*p*-value)	MCI vs. HC (*p*-value)
Demographics					
Female, %	50	30	42	0.6	0.6
Age, mean (sd)	71.8 (6.9)	74.0 (8.1)	69.8 (6.5)	0.50	0.20
Education, mean (sd)	13.3 (2.5)	13.4 (1.6)	13.2 (1.9)	0.90	0.90
Clinical characteristics					
RBANS.T, mean (sd)	64.3 (11.8)	84.9 (10.1)	107.6 (12.7)	**0.00**	**0.00**
MOCA.T, mean (sd)	16.8 (4.3)	22.7 (2.5)	27.0 (1.7)	**0.00**	**0.04**
Memory, mean (sd)	14.3 (7.7)	29.3 (6.7)	52.3 (6.9)	**0.00**	**0.00**
Language, mean (sd)	33.8 (15.0)	43.0 (13.1)	51.0 (6.1)	**0.00**	0.12
WM/Attention, mean (sd)	23.8 (18.6)	38.5 (11.5)	50.4 (8.1)	**0.00**	**0.00**
Visuoconstructional/EF, mean (sd)	18.3 (26.0)	43.3 (19.3)	49.1 (8.0)	**0.00**	**0.04**
Orientation, mean (sd)	−20.0 (51.1)	39.0 (17.1)	47.7 (16.4)	**0.00**	0.41
Structural characteristics					
Total gray matter (cm^3), mean (sd)	554.3 (55.5)	587.6 (37.2)	581.0 (59.2)	**0.05**	0.34

Significant differences (p < 0.05) between groups are highlighted in bold. RBANS.T, RBANS total score; MOCA.T, MOCA total score; WM, working memory; EF, executive function.

TABLE 2 | Mean and standard deviation of the speech characteristics per group and trial.

Speech measures	*Describe*-trial			*Match*-trial			*Describe and match*-trial		
Mean (sd)	**AD**	**MCI**	**HC**	**AD**	**MCI**	**HC**	**AD**	**MCI**	**HC**
N pauses	0.48 (0.17)	0.50 (0.14)	0.42 (0.17)	0.35 (0.10)	0.27 (0.12)	0.33 (0.14)	0.36 (0.16)	0.44 (0.11)	0.37 (0.08)
Gap/overlap ratio	10.35 (7.17)	8.88 (2.85)	6.30 (4.16)	10.43 (7.02)	9.42 (6.67)	5.25 (2.98)	6.12 (3.36)	8.95 (4.94)	5.43 (2.29)
Pause duration	0.94 (1.38)	0.82 (1.28)	0.60 (0.85)	0.85 (1.30)	1.01 (1.53)	0.78 (1.34)	0.64 (0.98)	0.92 (1.38)	0.66 (0.96)
Gap duration	1.13 (1.45)	1.04 (1.55)	0.69 (0.96)	1.36 (2.07)	1.18 (1.83)	0.78 (1.43)	1.23 (1.96)	1.00 (1.67)	0.69 (1.31)
Tov duration	0.17 (0.12)	0.24 (0.18)	0.34 (0.28)	0.22 (0.17)	0.25 (0.20)	0.29 (0.18	0.27 (0.23)	0.23 (0.20)	0.25 (0.19)
IPU duration	0.75 (0.56)	1.05 (0.76)	1.03 (0.80)	0.65 (0.52)	0.71 (0.54)	0.89 (0.67)	0.81 (0.65)	0.97 (0.74)	0.97 (0.68)
Turn duration	3.01 (3.78)	5.20 (5.96)	3.12 (3.68)	1.41 (1.82)	1.48 (2.16)	1.89 (3.04)	2.74 (4.44)	3.19 (4.77)	2.63 (4.31)
Speech rate	3.88 (1.76)	3.81 (1.61)	4.27 (1.93)	4.28 (2.07)	4.16 (2.03)	3.84 (2.09)	4.42 (1.73)	3.99 (1.51)	3.99 (2.03)

MCI, mild cognitive impairment; AD, Alzheimer's disease; HC, healthy controls; Describe-Trial: individuals with AD/MCI and HC are the directors, i.e., they describe the shapes and instruct where to place them; Match-Trial: individuals with AD/MCI and HC are the matchers, i.e., they place the pictures on their board following the caregiver's instructions; Describe and Match-Trial: both interlocutors describe the shapes and agree on where to place them. N, number of; tov, transition overlaps; IPU, interpausal unit.

TABLE 3 | Coefficients and 95% confidence intervals with marginal and conditional R^2 for the observed significant differences in speech features between groups ($p < 0.006$).

Groups	Speech features	Speech features * trials		R^2m; R^2c
	Speech rate	**Speech rate * trial**		
		Describe vs. *REF* trial	*Match* vs. *REF* trial	
AD vs. HC	–	−0.70 (1.05, −0.34)	–	0.04; 0.21
MCI vs. HC	–	–	–	
	Turn duration	**Turn duration * trial**		
		Describe vs. *REF* Trial	*Match* vs. *REF* Trial	
AD vs. HC	–	–	–	0.07; 0.12
MCI vs. HC	–	0.33 (0.05, 0.61)	−0.27 (−0.53, −0.01)	
	IPU duration	**IPU duration * trial**		
		Describe vs. *REF* Trial	*Match* vs. *REF* Trial	
AD vs. HC	−0.28 (−0.38, −0.06)	–	–	0.04; 0.10
MCI vs. HC	–	–	–	
	Tov duration	**Tov duration * trial**		
		Describe vs. *REF* Trial	*Match* vs. *REF* Trial	
AD vs. HC	–	−0.96 (−1.63, −0.24)	–	0.04; 0.10
MCI vs. HC	–	–	–	
	Gap/Overlap ratio	**Gap/Overlap ratio * trial**		
		Describe vs. *REF* Trial	*Match* vs. *REF* Trial	
AD vs. HC	–	–	–	0.18; 0.42
MCI vs. HC	0.54 (0.19, 0.89)	–	–	

The exact p-value is given in the text. AD, Alzheimer's disease; MCI, mild cognitive impairment; HC, healthy controls; Describe-Trial, individuals with AD/MCI and HC are the directors, i.e., they describe the shapes and instruct where to place them; Match-Trial, individuals with AD/MCI and HC are the matchers, i.e., they place the pictures on their board following the caregiver's instructions; REF-Trial (Describe and Match-Trial), both interlocutors describe the shapes and agree on where to place them; N, number of; tov, transition overlaps; IPU, interpausal unit.

R STG and R MFG; and interpausal unit with the L ITG, L IFG and L cereb. With regard to turn-taking organization, individuals with smaller R ITG volumes tended to produce longer gaps.

MCI and AD Classification Based on Temporal Speech Measures

Table 6 presents the cross-validated accuracy rates of the evolved classifiers for the test set based on the temporal speech features for the MCI-HC and AD-HC contrasts and for the multi-class classification of the three groups for the three trials separately. Best performances were achieved for the classifiers that were based on the speech measures extracted from Trial 1 and Trial 3 for the AD-HC contrast. Accuracy rates were moderate in the pairwise contrasts and slightly better for the AD-HC contrasts compared to the MCI-HC contrasts (e.g., 73.77 vs. 62.71 in Trial 1). Accuracy rates were similar across trials for the multi-class classification of the three groups (82.47% to 84.17%).

DISCUSSION

In this preliminary study exploring the cognitive and structural underpinnings of temporal speech characteristics in a collaborative referential task, we first show that MCI and mild-to-moderate AD are characterized by a general slowness of

TABLE 4 | Coefficients and 95% confidence intervals with marginal and conditional R^2 for the observed significant associations between the speech features and the cognitive domains (composite scores; $p < 0.006$).

Groups	Speech features	Speech features * trials		R^2m; R^2c
	Speech rate	**Speech rate * trial**		
		Describe vs. *REF* trial	*Match* vs. *REF* trial	
Memory	–	0.26 (0.05, 0.47)	–	0.04; 0.21
Visuoconst./EF	–	–	0.28 (0.08, 0.48)	
	Turn duration	**Turn duration * trial**		
		Describe vs. *REF* trial	*Match* vs. *REF* trial	
Memory	–	–	0.35 (0.16, 0.53)	0.06; 0.16
Visuoconst./EF	–	0.33 (0.17, 0.49)	0.18 (0.09, 0.14)	
WM/Attention	0.40 (0.20, 0.59)	−0.36 (−0.56, −0.16)	−0.40 (−0.59, −0.21)	
Orientation	–	−0.11 (−0.19, −0.02)	–	

The exact p-value is given in the text. Describe-Trial, individuals with AD/MCI and HC are the directors, i.e., they describe the shapes and instruct where to place them; Match-Trial, individuals with AD/MCI and HC are the matchers, i.e., they place the pictures on their board following the caregiver's instructions; REF-Trial (Describe and Match-Trial), both interlocutors describe the shapes and agree on where to place them; Visuoconst./EF, visuoconstruction/executive function; WM/Attention, working memory/attention; N, number of; tov, transition overlaps; IPU, interpausal unit.

TABLE 5 | Coefficients and 95% confidence intervals with marginal and conditional R^2 for the observed significant associations between the speech features and regional volumes in the fully adjusted models (per hemisphere; $p < 0.003$).

ROIs	Speech features	Speech features * trial		R^2m; R^2c
	Speech rate	**Speech rate * trial**		
		Describe vs. REF trial	*Match vs. REF trial*	
L MTG	–	–	−0.32 (−0.51, −0.14)	0.04; 0.22
R STG	–	–	0.30 (0.13, 0.47)	0.04; 0.22
R MTG	–	–	−0.43 (−0.60, −0.25)	
	Turn duration	**Turn duration * trial**		
		Describe vs. REF trial	*Match vs. REF trial*	
L MTG	−0.37 (−0.54, −0.20)	–	0.55 (0.39, 0.70)	0.09; 0.13
R MFG	–	0.14 (0.02, 0.25)	0.16 (0.05, 0.28)	0.07; 0.13
R STG	–	−0.17 (−0.29, −0.05)	−0.13 (−0.23, −0.01)	
	Interpausal unit duration	**Interpausal unit duration * trial**		
		Describe vs. REF trial	*Match vs. REF trial*	
L MTG	–	0.16 (0.00, 0.23)	0.28 (0.21, 0.36)	0.04; 0.12
L ITG	0.15 (0.03, 0.27)	−0.11 (−0.16, −0.04)	−0.23 (−0.30, −0.16)	
R IFG	–	−0.07 (−0.13, −0.02)	−0.10 (−0.16, −0.00)	0.04; 0.11
	Gap duration	**Gap duration * trial**		
		Describe vs. REF trial	*Match vs. REF trial*	
R ITG	–	−0.02 (−0.3, −0.00)	−0.02 (−0.03, −0.00)	0.05; 0.18

The exact p-value is given in the text. L, left; R, right; Cereb, cerebellum; FFG, fusiform gyrus; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; MFG, middle frontal gyrus; MTG, middle temporal gyrus; Prec, precuneus; STG, superior temporal gyrus; Describe-Trial, individuals with AD/MCI and HC are the directors, i.e., they describe the shapes and instruct where to place them; Match-Trial, individuals with AD/MCI and HC are the matchers, i.e., they place the pictures on their board following the caregiver's instructions; REF-Trial (Describe and Match-Trial), both interlocutors describe the shapes and agree on where to place them; N, number of; tov, transition overlaps; IPU, interpausal unit.

TABLE 6 | Cross-validated accuracy rates of the evolved classifiers for the test set based on temporal speech features for the MCI-HC and AD-HC contrasts and for the multi-class classification of the three groups for the three trials separately.

	AD/HC % (SD)	MC/HC % (SD)	Multi-class % (SD)
Describe-trial	73.77 (9.65)	62.71 (4.78)	83.95 (3.32)
Match-trial	63.67 (3.43)	63.41 (5.91)	82.47 (2.20)
Describe and match-trial	70.79 (9.50)	62.50 (7.67)	84.17 (2.72)

AD, Alzheimer's disease; MCI, mild cognitive impairment; HC, healthy controls; Describe-Trial, individuals with AD/MCI and HC are the directors, i.e., they describe the shapes and instruct where to place them; Match-Trial, individuals with AD/MCI and HC are the matchers, i.e., they place the pictures on their board following the caregiver's instructions; Describe-Match-Trial, both interlocutors describe the shapes and agree on where to place them.

speech, attributed to slower speech rate and slower turn-taking, with shorter transition overlaps and a larger number of gaps than transition overlap at speaker changes. Individuals with AD also had shorter interpausal units and individuals with

MCI had longer turns. Our findings on speech rate, pauses and interpausal units corroborate other analyses of connected speech in MCI and AD (Singh et al., 2001; Hoffmann et al., 2010; Rousseaux et al., 2010; Gayraud et al., 2011; Ahmed et al., 2013; Pistono et al., 2016; De Looze et al., 2018). The temporal characteristics of turn-taking organization, with slower exchanges for MCI and AD, support the potential existence of underlying cognitive deficits related to speech planning difficulties. Gap durations were almost doubled in the AD group compared to the HC (1,230 vs. 690 ms). Given that it takes about 1,500 ms to plan a simple sentence (Griffin and Bock, 2000), it may be postulated that individuals with AD needed more time to simultaneously comprehend their interlocutor's utterances, plan their answers and anticipate turn-endings.

More specifically, our analyses revealed that slower speech rate and longer pause duration were indicative of lower verbal memory scores and lower volumes of superior and middle temporal gyri. Slower speech rate was also associated with lower visuoconstructional/executive function scores and longer pauses with lower orientation scores.

Our findings suggest that slower speech rate and longer pause duration may be indicative of underlying deficits in episodic memory, lexical, semantic and executive functioning processes. Within the frame of the referential task, they may reflect difficulties with picture naming and remembering the sequence in which the pictures are described or remembering preceding exchanges (Feyereisen et al., 2007; Ash et al., 2011). Longer pauses may also reflect the time needed for the speaker to organize their thoughts and to construct a sentence. The associations observed with the superior and middle temporal gyri further support this interpretation. These regions have been linked to semantic memory and retrieval (Pravatà et al., 2016; Leyton et al., 2019) and to be dependent on an individual's verbal working memory capacity (Deschamps et al., 2014). Our findings corroborate the associations observed in other studies between within-clause pauses and activation in the superior and middle temporal gyri bilaterally (Kircher et al., 2004) as well as between speech rate and the STG and the MTG.

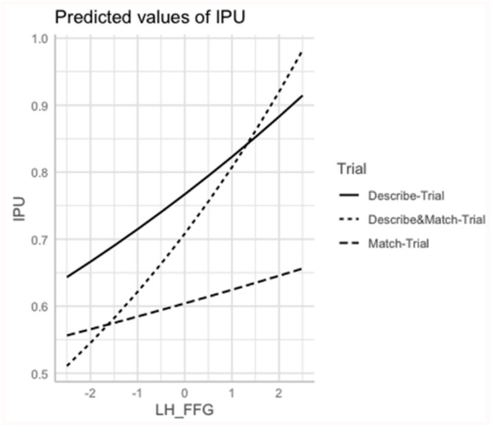


FIGURE 4 | Marginal estimates of Interpausal Unit (IPU) duration (log-transformed) as a function of the Left Fusiform Gyrus (L FFG) in the *Describe*, *Match*, and *Describe and Match* trials. *Describe*-Trial: individuals with AD/MCI and HC are the directors, i.e., they describe the shapes and instruct where to place them; *Match*-Trial: individuals with AD/MCI and HC are the matchers, i.e., they place the pictures on their board following the caregiver's instructions; *Describe and Match*-Trial: both interlocutors describe the shapes and agree on where to place them.

In addition, shorter interpausal units and shorter turns were associated with lower memory and working memory/attention scores. Shorter interpausal units were further related to lower orientation scores. Within the frame of the referential task, it may be hypothesized that these characteristics may reflect the production of shorter sentences of simpler syntactic and discourse structure and/or may be indicative of a narrower scope of speech planning (Swets et al., 2013; De Looze et al., 2018). With regards to the structural correlates, shorter interpausal units were associated with volume reductions in the right IFG, the left middle and inferior temporal gyri and left cerebellum. Associations were also observed between shorter turns and lower volumes of left and right middle MTG and right STG. The IFG is thought to support lexico-semantic retrieval processes and the parsing and planning of sentence and discourse-level linguistic information (Hagoort, 2005; Binder et al., 2009; Matsumoto et al., 2013; Hurley et al., 2015; Foti and Roberts, 2016). More generally, it has been linked to executive function, working memory and attention (Tops and Boksem, 2011; Zheng et al., 2014; Nissim et al., 2017). The inferior and middle frontal regions and the STG have been linked to speech planning processes and timing control. Furthermore, the cerebellum has been associated with speech and language control, timing, anticipation/prediction during language comprehension, verbal working memory and mental manipulation (Stoodley and Schmahmann, 2009; Marvel and Desmond, 2010; Murdoch, 2010; Mariën et al., 2014). In a previous study (De Looze et al., 2018), using data from the whole cohort (N = 80), we showed that the same temporal speech features in overt sentence reading were associated with reduced working memory/attention and language scores. We suggested that the temporal speech features may not only be reflective of lexico-semantic deficits but also of speech production planning difficulties, potentially stemming from reduced working memory capacity and attention deficits, specifically in the context of increased cognitive-linguistic demand. Several studies have provided evidence that the scope of speech production planning (i.e., how far ahead speakers plan an upcoming utterance) varies both as a function of speaker-specific verbal working memory capacity and cognitive-linguistic demands (Rochon et al., 2000; Swets et al., 2007; Petrone et al., 2011). We postulate that the size of interpausal units and turns may result from reduced working memory capacity in a highly cognitively demanding task, underlying speech production planning difficulties with reduced scope of planning (De Looze et al., 2018). The associations observed with the right hemisphere for several of these regions may be reflective of the nature of the referential task, also relying on visuoconstructional and visuospatial skills when describing the geometrical shapes and when ordering and placing the pictures on the board (Baddeley, 2000).

Finally, our exploratory analyses showed moderate accuracy rates for the speech-based classifiers in the pairwise contrasts, with higher performance for the AD-HC contrast (74%) compared to the MCI-HC contrast (63%). The accuracy rates for the multi-class classification of the three groups (84%) were in line with other studies also using ensembles of acoustic features derived from picture-description tasks, interviews or a combination of different speech tasks (Singh et al., 2001; Roark et al., 2011; Jarrold et al., 2014; Meilán et al., 2014; Dodge et al., 2015; König et al., 2015; López-de-Ipiña et al., 2015; Asgari et al., 2017; Tóth et al., 2018; Gosztolya et al., 2019; O'Malley et al., 2020). Using a combination of linear and nonlinear acoustic features extracted from spontaneous speech samples, López-de-Ipiña et al. (2015) reported 87% accuracy for the discrimination of AD. Similar features extracted from several short cognitive tasks were also used for the classification of MCI and AD, reaching accuracies of 79% and 87% for the MCI-HC and AD-HC contrasts respectively (König et al., 2015). Using a set of acoustic features extracted from longitudinally collected biographic interviews and cognitive tests, Weiner et al. (2016) achieved a classification accuracy of 86% between HCs, individuals with aging-associated cognitive decline and individuals with AD. Other studies (Jarrold et al., 2014; Gosztolya et al., 2019) have combined acoustic and lexical or linguistic features derived from spontaneous speech and reported an accuracy of 86–88% for the AD-HC contrast and 80% for the MCI-HC contrast.

These findings together support the discriminative power of speech-based approaches and their clinical relevance as a diagnostic tool component for the assessment and monitoring of cognitive deficits in ageing. The advantage with speech-based approaches is that they are less computationally demanding, they can be fully automated, they are non-invasive, time and cost-effective and are easy to administer. For example, speech changes could be recorded and monitored using a mobile phone. Anonymized data could be sent and processed to the cloud and feedback about an individual's cognitive functioning based on their speech characteristics, could be displayed and easily interpreted by a health professional *via* a web interface. Combining automated speech/language-based metrics with neuroimaging markers, neuropsychological scores and other behavioral measures, may assist health professionals in detecting and characterizing the course of cognitive decline in ageing and in defining an effective course of treatment and setting in place pertinent intervention strategies. These technologies may be of particular relevance in the context of stratification and screening procedures in overcrowded health services by providing some early insights (pending more in-depth clinical assessments) of an individual's cognitive function and potential underlying structural changes.

A number of limitations need to be highlighted. First the sample size of this study was small, and restricted to a specific age, education and cognitive functioning groups, which limits the generalizability of our results. Second, we opted for a Region-of-interest (ROI) based approach which may have left out some existing associations not investigated in this study. This approach was chosen to exploit the richness of the repeated measures collected per individual through linear mixed-effect modelling. Finally, it is not possible from the present observational study to infer any direction of causality and the interpretations in this manuscript provided can but only be speculative, although supported by accumulated evidence stemming from an extensive literature review.

The novelty of this study lies in the investigation of the association between temporal speech parameters, cognitive domains and brain regional volumes in a collaborative referential task. Our study explores for the first time the use of automatically extracted conversational-based features as input of EAs for the classification of MCI and AD while, at the same time, provides a thorough description of the cognitive and structural correlates of these features, with the modest intention of bringing clinical evidence of the relevance of these behavioral measures for the assessment and monitoring of MCI and AD.

CONCLUSION

Our study suggests that the temporal characteristics of speech in a collaborative referential task may reflect underlying cognitive deficits and structural volume reductions in healthy ageing, MCI and AD. The implementation of conversational speech-based technologies in clinical and community settings may represent a sensitive measure for the early assessment and longitudinal monitoring of cognitive-linguistic deficits and underlying structural changes in ageing.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not publicly available (following GPDR guidelines). Requests from research groups to access the datasets should be directed to deloozec@tcd.ie.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the St. James's Hospital Ethics and Medical Research Committee. Signed informed consent was obtained from all respondents prior to participation. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CD: drafting and stat analyses. AD: stat analyses and review. LC: data collection. AV: data annotation and review. RC, BL, and RR: input for data collection/analyses and review. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by a Centre for Ageing Research and Development in Ireland (CARDI) Leadership Fellowship (grant number 13533).

REFERENCES

Ahmed, O., Benois-Pineau, J., Allard, M., Catheline, G., and Amar, C. B. (2017). Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning. *Neurocomputing* 220, 98–110. doi: 10.1016/j.neucom.2016.08.041

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain Lang.* 136, 3727–3737. doi: 10.1093/brain/awt269

Albert, M. S., Steven, T. D., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 270–279. doi: 10.1016/j.jalz.2011.03.008

Asgari, M., Kaye, J., and Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimers Dement.* 3, 219–228. doi: 10.1016/j.trci.2017.01.006

Ash, S., McMillan, C., Gross, R. G., Cook, P., Morgan, B., Boller, A., et al. (2011). The organization of narrative discourse in Lewy body spectrum disorder. *Brain Lang.* 119, 30–41. doi: 10.1016/j.bandl.2011.05.006

Ash, S., Xie, S. X., Gross, R. G., Dreyfuss, M., Boller, A., Camp, E., et al. (2012). The organization and anatomy of narrative comprehension and expression in Lewy body spectrum disorders. *Neuropsychology* 26, 368–384. doi: 10.1037/a0027115

Bögels, S., Magyari, L., and Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Sci. Rep.* 5:12881. doi: 10.1038/srep12881

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends Cogn. Sci.* 4, 417–423. doi: 10.1016/s1364-6613(00)01538-2

Baldo, J. V., Arévalo, A., Patterson, J. P., and Dronkers, N. F. (2013). Gray and white matter correlates of picture naming: evidence from a voxel-based lesion analysis of the Boston Naming Test. *Cortex* 49, 658–667. doi: 10.1016/j.cortex.2012.03.001

Baldo, J. V., and Dronkers, N. F. (2006). The role of inferior parietal and inferior frontal cortex in working memory. *Neuropsychology* 20, 529–538. doi: 10.1037/0894-4105.20.5.529

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv* [Preprint]. doi: 10.18637/jss.v067.i01

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* 19, 2767–2796. doi: 10.1093/cercor/bhp055

Boersma, P., and Weenink, D. (2016). *Praat: Doing Phonetics by Computer [Computer program]. (Version 6.0. 23).* The Netherlands: Amsterdam.

Bourguignon, N. J. (2014). A rostro-caudal axis for language in the frontal lobe: the role of executive control in speech production. *Neurosci. Biobehav. Rev.* 47, 431–444. doi: 10.1016/j.neubiorev.2014.09.008

Brambati, S. M., Myers, D., Wilson, A., Rankin, K., Allison, S. C., Rosen, H. J., et al. (2006). The anatomy of categoryspecific object naming in neurodegenerative diseases. *J. Cogn. Neurosci.* 18, 1644–1653. doi: 10.1162/jocn.2006.18.10.1644

Caramelli, P., Lessa Mansur, L., and Nitrini, R. (1998). "Language and communication disorders in dementia of the Alzheimer type", *Handbook of Neurolinguistics*, eds B. Stemmer and H. A. Whitaker (San Diego, CA: Academic Press), 463–473.

Carlomagno, S., Santoro, A., Menditti, A., Pandolfi, M., and Marini, A. (2005). Referential communication in Alzheimer's type dementia. *Cortex* 41, 520–534. doi: 10.1016/s0010-9452(08)70192-8

Chan, D., Fox, N. C., Scahill, R. I., Crum, W. R., Whitwell, J. L., Leschziner, G., et al. (2001). Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Ann. Neurol.* 49, 433–442. doi: 10.1002/ana.92

Chapman, S. B., Zientz, J., Weiner, M., Rosenberg, R., Frawley, W., and Burns, M. H. (2002). Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. *Alzheimers Dis. Assoc. Disord.* 16, 177–186. doi: 10.1097/00002093-200207000-00008

Christodoulou, J. A., Del Tufo, S. N., Lymberis, J., Saxler, P. K., Ghosh, S. S., Triantafyllou, C., et al. (2014). Brain bases of reading fluency in typical reading and impaired fluency in dyslexia. *PLoS One* 9:e100552. doi: 10.1371/journal.pone.0100552

Croot, K., Hodges, J. R., Xuereb, J., and Patterson, K. (2000). Phonological and articulatory impairment in Alzheimer's disease: a case series. *Brain Lang.* 75, 277–309. doi: 10.1006/brln.2000.2357

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage* 9, 179–194. doi: 10.1006/nimg.1998.0395

Davis, B. H., and Maclagan, M. (2009). Examining pauses in Alzheimer's discourse. *Am. J. Alzheimers Dis. Other Dement.* 24, 141–154. doi: 10.1177/1533317508328138

de Jong, N. H., and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behav. Res. Methods* 41, 385–390. doi: 10.3758/BRM.41.2.385

De Looze, C., Kelly, F., Crosby, L., Vourdanou, A., Coen, R. F., Walsh, C., et al. (2018). Changes in speech chunking in reading aloud is a marker of mild cognitive impairment and mild-to-moderate Alzheimer's disease. *Curr. Alzheimer Res.* 15, 828–847. doi: 10.2174/156720501566618040416 5017

Dehsarvi, A. (2018). Classification of resting-state fMRI using evolutionary algorithms: towards a brain imaging biomarker for Parkinson's disease. PhD Thesis. University of York. Available online at: http://etheses.whiterose.ac.uk/20884.

Dehsarvi, A., and Smith, S. L. (2018). "Classification of resting-state fMRI for olfactory dysfunction in Parkinson's disease using evolutionary algorithms," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Kyoto, Japan, 264–265.

Deschamps, I., Baum, S. R., and Gracco, V. L. (2014). On the role of the supramarginal gyrus in phonological processing and verbal working memory: evidence from rTMS studies. *Neuropsychologia* 53, 39–46. doi: 10.1016/j.neuropsychologia.2013.10.015

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* 53, 1–15. doi: 10.1016/j.neuroimage.2010.06.010

Dickerson, B. C., Goncharova, I., Sullivan, M. P., Forchetti, C., Wilson, R. S., Bennett, D. A., et al. (2001). MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease. *Neurobiol. Aging* 22, 747–754. doi: 10.1016/s0197-4580(01)00271-8

Dicks, E., Tijms, B. M., Ten Kate, M., Gouw, A. A., Benedictus, M. R., Teunissen, C. E., et al. (2018). Gray matter network measures are associated with cognitive decline in mild cognitive impairment. *Neurobiol. Aging* 61, 198–206. doi: 10.1016/j.neurobiolaging.2017.09.029

Dodge, H., Mattek, N., Gregor, M., Bowman, M., Seelye, A., Ybarra, O., et al. (2015). Social markers of mild cognitive impairment: proportion of word counts in free conversational speech. *Curr. Alzheimer Res.* 12, 513–519. doi: 10.2174/1567205012666150530201917

Drummond, C., Coutinho, G., Fonseca, R. P., Assunção, N., Teldeschi, A., de Oliveira-Souza, R., et al. (2015). Fernanda Tovar-Moll and Paulo Mattos deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. *Front. Aging Neurosci.* 7:96. doi: 10.3389/fnagi.2015.00096

Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., et al. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol.* 6, 734–746. doi: 10.1016/S1474-4422(07)70178-3

Duff, M. C., Gallegos, D. R., Cohen, N. J., and Tranel, D. (2013). Learning in Alzheimer's disease is facilitated by social interaction. *J. Comp. Neurol.* 521, 4356–4369. doi: 10.1002/cne.23433

Ferreira, F., and Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *J. Mem. Lang.* 46, 57–84. doi: 10.1006/jmla.2001.2797

Feyereisen, P., Berrewaerts, J., and Hupet, M. (2007). Pragmatic skills in the early stages of Alzheimer's disease: an analysis by means of a referential communication task. *Int. J. Lang. Commun. Disord.* 42, 1–17. doi: 10.1080/13682820600624216

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355. doi: 10.1016/s0896-6273(02)00569-x

Forbes-McKay, K., Shanks, M. F., and Venneri, A. (2013). Profiling spontaneous speech decline in Alzheimer's disease: a longitudinal study. *Acta Neuropsychiatr.* 25, 320–327. doi: 10.1017/neu.2013.16

Foti, D., and Roberts, F. (2016). The neural dynamics of speech perception: dissociable networks for processing linguistic content and monitoring speaker turn-taking. *Brain Lang.* 157, 63–71. doi: 10.1016/j.bandl.2016.05.001

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49, 407–422. doi: 10.3233/JAD-150520

Gayraud, F., Lee, H.-R., and Barkat-Defradas, M. (2011). Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects. *Clin. Linguist. Phon.* 25, 198–209. doi: 10.3109/02699206.2010.521612

Goldman Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech.* London: Academic Press.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., and Hoffmann, I. (2019). Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Comput. Speech Lang.* 53, 181–197. doi: 10.1016/j.csl.2018.07.007

Griffin, Z. M., and Bock, K. (2000). What the eyes say about speaking. *Psychol. Sci.* 11, 274–279. doi: 10.1111/1467-9280.00255

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9, 416–423. doi: 10.1016/j.tics.2005.07.004

Hampson, M., Tokoglu, F., Sun, Z., Schafer, R. J., Skudlarski, P., Gore, J. C., et al. (2006). Connectivity-behavior analysis reveals that functional connectivity between left BA39 and Broca's area varies with reading ability. *NeuroImage* 31, 513–519. doi: 10.1016/j.neuroimage.2005.12.040

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., et al. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *NeuroImage* 32, 180–194. doi: 10.1016/j.neuroimage.2006.02.051

Helder, A., van den Broek, P., Karlsson, J., and Van Leijenhorst, L. (2017). Neural correlates of coherence-break detection during reading of narratives. *Sci. Stud. Reading* 21, 463–479. doi: 10.1080/10888438.2017.1332065

Hirshorn, E. A., Dye, M. W. G., Hauser, P. C., Supalla, T. R., and Bavelier, D. (2014). Neural networks mediating sentence reading in the deaf. *Front. Hum. Neurosci.* 8:394. doi: 10.3389/fnhum.2014.00394

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., and Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *Int. J. Speech Lang. Pathol.* 12, 29–34. doi: 10.3109/17549500903137256

Hurley, R. S., Bonakdarpour, B., Wang, X., and Mesulam, M. M. (2015). Asymmetric connectivity between the anterior temporal lobe and the language network. *J. Cogn. Neurosci.* 27, 464–473. doi: 10.1162/jocn_a_00722

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., et al. (2014). "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Baltimore, MD), 27–37.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van Der Kouwe, A., Gollub, R., et al. (2006). Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *NeuroImage* 30, 436–443. doi: 10.1016/j.neuroimage.2005.09.046

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement.* 1, 112–124. doi: 10.1016/j.dadm.2014.11.012

Killiany, R. J., Hyman, B. T., Gomez-Isla, T., Moss, M. B., Kikinis, R., Jolesz, F., et al. (2002). MRI measures of entorhinal cortex vs. hippocampus in preclinical AD. *Neurology* 58, 1188–1196. doi: 10.1212/wnl.58.8.1188

Kircher, T. T., Brammer, M. J., Levelt, W., Bartels, M., and McGuire, P. K. (2004). Pausing for thought: engagement of left temporal cortex during pauses in speech. *NeuroImage* 21, 84–90. doi: 10.1016/j.neuroimage.2003.09.041

Kirova, A.-M., Bays, R. B., and Lagalwar, S. (2015). Working memory and executive function decline across normal aging, mild cognitive impairment, and Alzheimer's disease. *Biomed Res. Int.* 2015:748212. doi: 10.1155/2015/748212

López-de-Ipiña, K., Solé-Casals, J., Eguiraun, H., Alonso, J. B., Travieso, C. M., Ezeiza, A., et al. (2015). Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: a fractal dimension approach. *Comput. Speech Lang.* 30, 43–60. doi: 10.1016/j.csl.2014.08.002

Laws, K. R., Duncan, A., and Gale, T. M. (2010). 'Normal' semantic-phonemic fluency discrepancy in Alzheimer's disease? A meta-analytic study. *Cortex* 46, 595–601. doi: 10.1016/j.cortex.2009.04.009

Lehéricy, S., Baulac, M., Chiras, J., Piérot, L., Martin, N., Pillon, B., et al. (1994). Amygdalohippocampal MR volume measurements in the early stages of Alzheimer disease. *Am. J. Neuroradiol.* 15, 929–937.

Leyton, C. E., Landin-Romero, R., Liang, C. T., Burrell, J. R., Kumfor, F., Hodges, J. R., et al. (2019). Correlates of anomia in non-semantic variants of primary progressive aphasia converge over time. *Cortex* 120, 201–211. doi: 10.1016/j.cortex.2019.06.008

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models.* Cary, NC: SAS Institute, Inc.

Magyari, L., Bastiaansen, M. C., de Ruiter, J. P., and Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *J. Cogn. Neurosci.* 26, 2530–2539. doi: 10.1162/jocn_a_00673

Mariën, P., Ackermann, H., Adamaszek, M., Barwood, C. H. S., Beaton, A., Desmond, J., et al. (2014). Consensus paper: language and the cerebellum: an ongoing enigma. *Cerebellum* 13, 386–410. doi: 10.1007/s12311-013-0540-5

Marvel, C. L., and Desmond, J. E. (2010). Functional topography of the cerebellum in verbal working memory. *Neuropsychol. Rev.* 20, 271–279. doi: 10.1007/s11065-010-9137-7

Matsumoto, K., Kircher, T., Stokes, P., Brammer, M. J., LIddle, P., and McGuire, P. K. (2013). Frequency and neural correlates of pauses in patients with formal thought disorder. *Front. Psychiatry* 4:127. doi: 10.3389/fpsyt.2013.00127

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R. Jr., Kawas, C. H., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 263–269. doi: 10.1016/j.jalz.2011.03.005

Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement. Geriatr. Cogn. Disord.* 37, 327–334. doi: 10.1159/000356726

Miller, J. F. (2020). Cartesian genetic programming: its status and future. *Genet. Prog. Evolvable Mach.* 21, 129–168. doi: 10.1007/s10710-019-09360-6

Mirheidari, B., Blackburn, D., O'Malley, R., Venneri, A., Walker, T., Reuber, M., et al. (2020). "Improving cognitive impairment classification by generative neural network-based feature augmentation," in *Proceedings of the Interspeech*, Shanghai, China, 2527–2531.

Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. *J. Clin. Exp. Neuropsychol.* 40, 917–939. doi: 10.1080/13803395.2018.1446513

Muhamed, S. A., Newby, R., Smith, S. L., Alty, J. E., Jamieson, S., and Kempster, P. (2018). "Objective evaluation of bradykinesia in Parkinson's disease using evolutionary algorithms," in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies—Volume 4: BIOSIGNALS* (Funchal, Madeira, Portugal), 63–69.

Murdoch, B. E. (2010). The cerebellum and language: historical perspective and review. *Cortex* 46, 858–868. doi: 10.1016/j.cortex.2009.07.018

Murdoch, B. E., Chenery, H. J., Wilks, V., and Boyle, R. S. (1987). Language disorders in dementia of the Alzheimer type. *Brain Lang.* 31, 122–137. doi: 10.1016/0093-934x(87)90064-2

Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* 4, 133–142. doi: 10.1111/j.2041-210x.2012.00261.x

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Newman, S. D., Malaia, E., Seo, R., and Cheng, H. (2013). The effect of individual differences in working memory capacity on sentence comprehension: an fMRI study. *Brain Topogr.* 26, 458–467. doi: 10.1007/s10548-012-0264-8

Nissim, N. R., O'Shea, A. M., Bryant, V., Porges, E. C., Cohen, R., and Woods, A. J. (2017). Frontal structural neural correlates of working memory performance in older adults. *Front. Aging Neurosci.* 8:328. doi: 10.3389/fnagi.2016.00328

O'Malley, R. P. D., Mirheidari, B., Harkness, K., Reuber, M., Venneri, A., Walker, T., et al. (2020). A fully automated cognitive screening tool based on assessment of speech and language. *J. Neurol. Neurosurg. Psychiatry* 92, 12–15. doi: 10.1136/jnnp-2019-322517

Petrone, C., Fuchs, S., and Krivokapić, J. (2011). "Consequences of working memory differences and phrasal length on pause duration and fundamental frequency," in *Proceedings of the 9th International Seminar on Speech Production (ISSP)*, Montreal, Canada, 393–400.

Picardi, C., Cosgrove, J., Smith, S. L., Jamieson, S., and Alty, J. E. (2017). "Objective assessment of cognitive impairment in Parkinson's disease using evolutionary algorithm," in *European Conference on the Applications of Evolutionary Computation* (Cham, Springer), 109–124.

Pistono, A., Jucla, M., Barbeau, E. J., Saint-Aubert, L., Lemesle, B., Calvet, B., et al. (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *J. Alzheimers Dis.* 50, 687–698. doi: 10.3233/JAD-150408

Prat, C. S., Keller, T. A., and Adam Just, M. (2007). Individual differences in sentence comprehension: a functional magnetic resonance imaging investigation of syntactic and lexical processing demands. *J. Cogn. Neurosci.* 19, 1950–1963. doi: 10.1162/jocn.2007.19.12.1950

Pravatà, E., Tavernier, J., Parker, R., Vavro, H., Mintzer, J. E., and Spampinato, M. V. (2016). The neural correlates of anomia in the conversion from mild cognitive impairment to Alzheimer's disease. *Neuroradiology* 58, 59–67. doi: 10.1038/s41594-020-00556-4

Randolph, C., Tierney, M. C., Mohr, E., and Chase, T. N. (1998). The repeatable battery for the assessment of neuropsychological status (RBANS): preliminary clinical validity. *J. Clin. Exp. Neuropsychol.* 20, 310–319. doi: 10.1076/jcen.20.3.310.823

R Core Team. (2015). *R: A Language and Environment for Statistical Computing.* Available online at: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing.

Riest, C., Jorschick, A. B., and de Ruiter, J. P. (2015). Anticipation in turn-taking: mechanisms and information sources. *Front. Psychol.* 6:89. doi: 10.3389/fpsyg.2015.00089

Ripich, D. N., Vertes, D., Whitehouse, P., Fulton, S., and Ekelman, B. (1991). Turn-taking and speech act patterns in the discourse of senile dementia of the Alzheimer's type patients. *Brain Lang.* 40, 330–343. doi: 10.1016/0093-934x(91)90133-l

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans. Audio Speech Lang. Process.* 19, 2081–2090. doi: 10.1109/TASL.2011.2112351

Rochon, E., Waters, G. S., and Caplan, D. (2000). The relationship between measures of working memory and sentence comprehension in patients with Alzheimer's disease. *J. Speech Lang. Hear. Res.* 43, 395–413. doi: 10.1044/jslhr.4302.395

Rousseaux, M., Sève, A., Vallet, M., Pasquier, F., and Mackowiak-Cordoliani, M. A. (2010). An analysis of communication in conversation in patients with dementia. *Neuropsychologia* 48, 3884–3890. doi: 10.1016/j.neuropsychologia.2010.09.026

Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). "A simplest systematics for the organization of turn taking for conversation," in *Studies in the Organization of Conversational Interaction*, ed Jim Schenkein (Academic Press), 7–55.

Sajjadi, S. A., Patterson, K., Tomek, M., and Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs. Alzheimer's disease. *Aphasiology* 26, 847–866. doi: 10.1080/02687038.2012.654933

Sanderman, A. A., and Collier, R. (1995). "Prosodic phrasing at the sentence level," in *Producing Speech: Contemporary Issues: for Katherine Safford Harris*, eds F. Bell-Berti, L. J. Raphael (New York: AIP Press), 321–332.

Singh, S., Bucks, R. S., and Cuerden, J. M. (2001). Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech. *Aphasiology* 15, 571–583. doi: 10.1080/02687040143000041

Sohn, J., Kim, N., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Process. Lett.* 6, 1–3.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 280–292. doi: 10.1016/j.jalz.2011.03.003

Stoodley, C. J., and Schmahmann, J. D. (2009). Functional topography in the human cerebellum: a meta-analysis of neuroimaging studies. *NeuroImage* 44, 489–501. doi: 10.1016/j.neuroimage.2008.08.039

Swets, B., Desmet, T., Hambrick, D. Z., and Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: a psychometric approach. *J. Exp. Psychol. Gen.* 136, 64–81. doi: 10.1037/0096-3445.136.1.64

Swets, B., Jacovina, M. E., and Gerrig, R. J. (2013). Effects of conversational pressures on speech planning. *Dis. Process.* 50, 23–51. doi: 10.1080/0163853x.2012.727719

Swets, B., Jacovina, M. E., and Gerrig, R. J. (2014). Individual differences in the scope of speech planning: evidence from eye-movements. *Lang. Cogn.* 6, 12–44. doi: 10.1017/langcog.2013.5

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Front. Aging Neurosci.* 7:195. doi: 10.3389/fnagi.2015.00195

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* 15, 130–138. doi: 10.2174/1567205014666171121114930

Taler, V., and Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *J. Clin. Exp. Neuropsychol.* 30, 501–556. doi: 10.1080/13803390701550128

Tondelli, M., Wilcock, G. K., Nichelli, P., De Jager, C. A., Jenkinson, M., and Zamboni, G. (2012). Structural MRI changes detectable up to ten years before clinical Alzheimer's disease. *Neurobiol. Aging* 33, 825.e25–825.e36. doi: 10.1016/j.neurobiolaging.2011.05.018

Tops, M., and Boksem, M. A. S. (2011). A potential role of the inferior frontal gyrus and anterior insula in cognitive control, brain rhythms, and event-related potentials. *Front. Psychol.* 2:330. doi: 10.3389/fpsyg.2011.00330

Turner, A. J., and Miller, J. F. (2015). Introducing a cross platform open source cartesian genetic programming library. *Genet. Prog. Evolvable Mach.* 16, 83–91. doi: 10.1007/s10710-014-9233-1

Verfaillie, S. C., Slot, R. E. R., Dicks, E., Prins, N. D., Overbeek, J. M., Teunissen, C. E., et al. (2018). A more randomly organized gray matter network is associated with deteriorating language and global cognition in individuals with subjective cognitive decline. *Hum. Brain Mapp.* 39, 3143–3151. doi: 10.1002/hbm.24065

Wang, W.-Y., Yu, J.-T., Liu, Y., Yin, R.-H., Wang, H.-F., Wang, J., et al. (2015). Voxel-based meta-analysis of gray matter changes in Alzheimer's disease. *Transl. Neurodegener.* 4:6. doi: 10.1186/s40035-015-0027-z

Weiner, J., Herff, C., and Schultz, T. (2016). "Speech-based detection of Alzheimer's disease in conversational german," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, San Francisco, California, USA, 1938–1942.

Weiner, M. F., Neubecker, K. E., Bret, M. E., and Hynan, L. S. (2008). Language in Alzheimer's disease. *J. Clin. Psychiatry* 69, 1223–1227. doi: 10.4088/jcp.v69n0804

Xu, J., Kemeny, S., Park, G., Frattali, C., and Braun, A. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage* 25, 1002–1015. doi: 10.1016/j.neuroimage.2004.12.013

Zheng, D., Sun, H., Dong, X., Liu, B., Xu, Y., Chen, S., et al. (2014). Executive dysfunction and gray matter atrophy in amnestic mild cognitive impairment. *Neurobiol. Aging* 35, 548–555. doi: 10.1016/j.neurobiolaging.2013.09.007

Conflict of Interest: The 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.

frontiers in Aging Neuroscience

METHODS
published: 27 April 2021
doi: 10.3389/fnagi.2021.635945

Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech

Aparna Balagopalan[1,2,3]*, Benjamin Eyre*[1]*, Jessica Robin*[1]*, Frank Rudzicz*[2,3,4] *and Jekaterina Novikova*[1]*

[1] *Winterlight Labs Inc., Toronto, ON, Canada,* [2] *Department of Computer Science, University of Toronto, Toronto, ON, Canada,* [3] *Vector Institute for Artificial Intelligence, Toronto, ON, Canada,* [4] *Unity Health Toronto, Toronto, ON, Canada*

Introduction: Research related to the automatic detection of Alzheimer's disease (AD) is important, given the high prevalence of AD and the high cost of traditional diagnostic methods. Since AD significantly affects the content and acoustics of spontaneous speech, natural language processing, and machine learning provide promising techniques for reliably detecting AD. There has been a recent proliferation of classification models for AD, but these vary in the datasets used, model types and training and testing paradigms. In this study, we compare and contrast the performance of two common approaches for automatic AD detection from speech on the same, well-matched dataset, to determine the advantages of using domain knowledge vs. pre-trained transfer models.

Methods: Audio recordings and corresponding manually-transcribed speech transcripts of a picture description task administered to 156 demographically matched older adults, 78 with Alzheimer's Disease (AD) and 78 cognitively intact (healthy) were classified using machine learning and natural language processing as "AD" or "non-AD." The audio was acoustically-enhanced, and post-processed to improve quality of the speech recording as well control for variation caused by recording conditions. Two approaches were used for classification of these speech samples: (1) using domain knowledge: extracting an extensive set of clinically relevant linguistic and acoustic features derived from speech and transcripts based on prior literature, and (2) using transfer-learning and leveraging large pre-trained machine learning models: using transcript-representations that are automatically derived from state-of-the-art pre-trained language models, by fine-tuning Bidirectional Encoder Representations from Transformer (BERT)-based sequence classification models.

Results: We compared the utility of speech transcript representations obtained from recent natural language processing models (i.e., BERT) to more clinically-interpretable language feature-based methods. Both the feature-based approaches and fine-tuned BERT models significantly outperformed the baseline linguistic model using a small set of linguistic features, demonstrating the importance of extensive linguistic information for detecting cognitive impairments relating to AD. We observed that fine-tuned BERT

Edited by:
Saturnino Luz,
University of Edinburgh, United Kingdom

Reviewed by:
Kewei Chen,
Banner Alzheimer's Institute, United States
Juan José García Meilán,
University of Salamanca, Spain

***Correspondence:**
Jekaterina Novikova
jekaterina@winterlightlabs.com

Received: *30 November 2020*
Accepted: *24 March 2021*
Published: *27 April 2021*

Citation:
Balagopalan A, Eyre B, Robin J, Rudzicz F and Novikova J (2021) Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech.
Front. Aging Neurosci. 13:635945.
doi: 10.3389/fnagi.2021.635945

models numerically outperformed feature-based approaches on the AD detection task, but the difference was not statistically significant. Our main contribution is the observation that when tested on the same, demographically balanced dataset and tested on independent, unseen data, both domain knowledge and pretrained linguistic models have good predictive performance for detecting AD based on speech. It is notable that linguistic information alone is capable of achieving comparable, and even numerically better, performance than models including both acoustic and linguistic features here. We also try to shed light on the inner workings of the more black-box natural language processing model by performing an interpretability analysis, and find that attention weights reveal interesting patterns such as higher attribution to more important information content units in the picture description task, as well as pauses and filler words.

Conclusion: This approach supports the value of well-performing machine learning and linguistically-focussed processing techniques to detect AD from speech and highlights the need to compare model performance on carefully balanced datasets, using consistent same training parameters and independent test datasets in order to determine the best performing predictive model.

Keywords: Alzheimer's disease, dementia detection, MMSE regression, BERT, feature engineering, transfer learning

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disease that causes problems with memory, thinking, and behavior. AD affects over 40 million people worldwide with high costs of acute and long-term care (Prince et al., 2016). Current forms of diagnosis are both time consuming and expensive (Prabhakaran et al., 2018), which might explain why almost half of those living with AD do not receive a timely diagnosis (Jammeh et al., 2018).

Studies have shown that valuable clinical information indicative of cognition can be obtained from spontaneous speech elicited using pictures (Goodglass et al., 2001). Studies have capitalized on this clinical observation, using speech analysis, natural language processing (NLP), and machine learning (ML) to distinguish between speech from healthy and cognitively impaired participants in datasets including semi-structured speech tasks such as picture description. Some of the first papers on this topic reported ML methods for automatic AD-detection using speech datasets achieving high classification performance (between 82 and 93% accuracy) (König et al., 2015; Fraser et al., 2016; Noorian et al., 2017; Karlekar et al., 2018; Zhu et al., 2018; Gosztolya et al., 2019). These models serve as quick, objective, and non-invasive assessments of an individual's cognitive status which could be developed into more accessible tools to facilitate clinical screening and diagnosis. Since these initial reports, there has been a proliferation of studies reporting classification models for AD based on speech, as described by recent reviews and meta-analyses (Slegers et al., 2018; de la Fuente Garcia et al., 2020; Petti et al., 2020; Pulido et al., 2020), but the field still lacks validation of predictive models on publicly-available, balanced, and standardized benchmark datasets.

The existing studies that have addressed differences between AD and non-AD speech and worked on developing speech-based AD biomarkers, are often descriptive rather than predictive. Thus, they often overlook common biases in evaluations of AD detection methods, such as repeated occurrences of speech from the same participant, variations in audio quality of speech samples, and imbalances of gender and age distribution in the used datasets, as noted in the systematic reviews and meta-analyses published on this topic (Slegers et al., 2018; Chen et al., 2020; Petti et al., 2020). As such, the existing ML models may be prone to the biases introduced in available data. In addition, the performance of the previously developed predictive AD-detection models has been evaluated using either random train/test split or a cross-validation technique, which may result in artificially increased reported performance of ML models (i.e., overfitting) as compared to their evaluation on a held out unseen dataset (more details on evaluation techniques are provided in the section 2.3.1.2), especially when it comes to smaller and unbalanced datasets (Johnson et al., 2018). Due to these reasons, it's difficult to compare model performance across papers and datasets, since they are rarely matched in terms of data and model characteristics.

To overcome the problem of bias and overfitting and introduce a common dataset to compare model performance, the ADReSS challenge (Luz et al., 2020) was introduced in 2020, in which the organizers provided an age/sex-matched balanced speech dataset, which consisted of speech from AD and non-AD participants describing a picture. The challenge consisted of two key tasks: (1) Speech classification task: classifying speech as

AD or non-AD. (2) Neuropsychological score regression task: predicting Mini-Mental State Examination (MMSE) (Cockrell and Folstein, 2002) scores from speech. The organizers restricted access to the test dataset to make it completely unseen for participants to ensure the fair evaluation of models' performance. The work presented in this paper is focused entirely on this new balanced dataset and follows the ADReSS challenge's evaluation process. As such, the models presented in this paper are more generalizable to unseen data than those developed in the previously discussed studies.

In this work, we develop ML models to detect AD from speech using picture description data of the demographically-matched ADReSS Challenge speech dataset (Luz et al., 2020), and compare the following training regimes and input representations to detect AD:

1. **Using domain knowledge**: with this approach, we extract clinically relevant linguistic features from transcripts of speech, and acoustic features from corresponding audio files for binary AD vs. non-AD classification and MMSE score regression. The features extracted are informed by previous clinical and ML research in the space of cognitive impairment detection (Fraser et al., 2016).
2. **Using transfer learning**: with this approach, we fine-tune pre-trained BERT (Devlin et al., 2019) text classification models at transcript-level.

We describe below the details of each approach.

1.1. Domain Knowledge-Based Approach

The overwhelming majority of NLP and ML approaches on AD detection from speech are still based on hand-crafted engineering of clinically-relevant features (de la Fuente Garcia et al., 2020). Previous work that focused on automatic AD detection from speech uses certain acoustic features (such as zero-crossing rate, Mel-frequency cepstral coefficients etc.) and linguistic features (such as proportions of various parts-of-speech (POS) tags (Orimaye et al., 2015; Fraser et al., 2016; Noorian et al., 2017), etc.) from speech transcripts. Fraser et al. (2016) extracted 370 linguistic and acoustic features from picture descriptions in the DementiaBank dataset, and obtained an AD detection accuracy of 82% at transcript-level. Fraser et al.'s model was evaluated using cross-validation. More recent studies showed the addition of normative data helped increase accuracy up to 93%, when evaluated using a random train/test split (Noorian et al., 2017; Balagopalan et al., 2018). Yancheva et al. (2015) showed ML models are capable of predicting the MMSE scores from features of speech elicited via picture descriptions, with mean absolute error of 2.91-3.83.

Detecting AD or predicting MMSE scores with pre-engineered features of speech and thereby infusing domain knowledge into the task has several advantages, such as more interpretable model decisions, the possibility to represent speech in different modalities (both acoustic and linguistic), and potentially lower computational resource requirements when paired with conventional ML models. However, there are also a few disadvantages, e.g., a feature engineering process is very expensive and time-consuming, it requires clinical expertise, is prone to biases in data, and carries the risk of missing highly relevant features.

1.2. Transfer Learning-Based Approach

In the recent years, transfer learning, or in other words, utilizing language representations from huge pre-trained neural models that learn robust representations for text, has become ubiquitous in NLP (Young et al., 2018). One of the most popular transfer learning models is BERT (Devlin et al., 2019), which trains "contextual embeddings" wherein a representation of a sentence (or transcript) is influenced by the context in which the words occur in sentences. This model offers enhanced parallelization and better modeling of long-range dependencies in text and as such, has achieved state-of-the-art performance on a variety of tasks in NLP. Previous research (Jawahar et al., 2019; Rogers et al., 2021) has suggested that it encodes language information (lexical, syntactic etc.) that is known to be important for performing complex natural language tasks, including AD detection from speech.

BERT uses powerful attention mechanisms to encode global dependencies between the input and output. This allows it to achieve state-of-the-art results on a suite of benchmarks (Devlin et al., 2019). Fine-tuning BERT for a few epochs can potentially attain good performance even on small datasets.

The transfer learning technique in general and BERT model specifically are promising approaches to apply to the task of AD detection from speech because such a technique eliminates the need of expensive and time-consuming feature engineering, mitigates the need of big training datasets, and potentially results in more generalizable models. However, the common critique is that BERT is pre-trained on the corpus of healthy language and as such is not usable for detecting AD. In addition, BERT is not directly interpretable, unlike feature-based models. Finally, the original version of the BERT model is only able to use text as input, thus eliminating the possibility to employ the acoustic modality of speech, when detecting AD. All these may be the reasons why BERT was not previously used for developing predictive models for AD detection, even though its performance on many other NLP tasks is exceptional.

1.3. Motivation and Contributions

Our motivation in this work is to benchmark a BERT training procedure on transcripts from a pathological speech dataset, and evaluate the effectiveness of high-level language representations from BERT in detecting AD. We are specifically interested in understanding whether BERT has a potential to outperform traditional widely used domain-knowledge based approaches given that it does not include acoustic features, and at the same time increase the generalizability of the predictive models.

To eliminate the biases of unbalanced data, we perform all our experiments on the carefully demographically-matched ADReSS dataset. To understand how well the presented models generalize to unseen data, we evaluate performance of the models using both cross-validation and testing on unseen held out dataset.

We find that the feature-based SVM model with RBF kernel outperforms all the other models, and performs on par with BERT, when evaluated using cross-validation. When

TABLE 1 | Basic characteristics of the patients in each group in the ADReSS challenge dataset are more balanced in comparison to DementiaBank.

Dataset			Class	
			AD	Non-AD
ADReSS	Train	Male	24	24
		Female	30	30
ADReSS	Test	Male	11	11
		Female	13	13
DementiaBank (Becker et al., 1994)	–	Male Female	125 197	83 146

TABLE 2 | ADReSS Training set from Luz et al. (2020): basic characteristics of the patients in each group (M, male; F, female).

	AD			Non-AD		
Age	M	F	MMSE (sd)	M	F	MMSE (sd)
[50, 55)	1	0	30.0 (n/a)	1	0	29.0 (n/a)
[55, 60)	5	4	16.3 (4.9)	5	4	29.0 (1.3)
[60, 65)	3	6	18.3 (6.1)	3	6	29.3 (1.3)
[65, 70)	6	10	16.9 (5.8)	6	10	29.1 (0.9)
[70, 75)	6	8	15.8 (4.5)	6	8	29.1 (0.8)
[75, 80)	3	2	17.2 (5.4)	3	2	28.8 (0.4)
Total	24	30	17.0 (5.5)	24	30	29.1 (1.0)

TABLE 3 | ADReSS test set from Luz et al. (2020): basic characteristics of the patients in each group (M, male; F, female).

	AD			Non-AD		
Age	M	F	MMSE (sd)	M	F	MMSE (sd)
[50, 55)	1	0	23.0 (n.a)	1	0	28.0 (n.a)
[55, 60)	2	2	18.7 (1.0)	2	2	28.5 (1.2)
[60, 65)	1	3	14.7 (3.7)	1	3	28.7 (0.9)
[65, 70)	3	4	23.2 (4.0)	3	4	29.4 (0.7)
[70, 75)	3	3	17.3 (6.9)	3	3	28.0 (2.4)
[75, 80)	1	1	21.5 (6.3)	1	1	30.0 (0.0)
Total	11	13	19.5 (5.3)	11	13	28.8 (1.5)

evaluation is performed on the unseen held out test data, the fine-tuned BERT text sequence classification models achieve the highest AD detection accuracy of 83.3%. This BERT model numerically, though not significantly, outperforms the SVM model that achieves 81.3% accuracy on the unseen test set. These results show that: (1) Extensive feature-based—i.e., containing linguistic information for various aspects of language such as semantics, syntax, and lexicon—classification models significantly outperforms the linguistic baseline provided in the challenge showing that feature engineering to capture various aspects of language such as semantics and syntax helps with reliable detection of AD from speech, (2) BERT proved to be a generalizable model comparable to feature-based ones that make use of domain knowledge via hand-crafted feature engineering as shown by its higher performance on the independent test set in our case, (3) linguistic-only information encoded in BERT is sufficient for the strong predictive performance of the AD detection models.

2. MATERIALS AND METHODS

2.1. ADReSS Dataset

Our data are derived from the ADReSS Challenge dataset (Luz et al., 2020), which consists of 156 speech recordings and associated transcripts from non-AD (N = 78) and AD (N = 78) English-speaking participants. Speech is elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia exam (Goodglass et al., 2001). Transcripts were annotated using the CHAT coding system (MacWhinney, 2000). In contrast to other speech datasets for AD detection such as DementiaBank's English Pitt Corpus (Becker et al., 1994), the ADReSS challenge dataset is carefully matched for age and gender in order to minimize risk of bias in the prediction tasks (**Tables 1–3**). Recordings were acoustically enhanced by the challenge organizers with stationary noise removal and audio volume normalization was applied across all speech segments to control for variation caused by recording conditions such as microphone placement (Luz et al., 2020). The speech dataset is divided into the train set and the unseen held out test set. MMSE (Cockrell and Folstein, 2002) scores are available for all but one of the participants in the train set.

2.2. Feature Extraction

The speech transcripts in the dataset are manually transcribed as per the CHAT protocol (MacWhinney, 2000), and include speech segments from both the participant and an investigator. We only use the portion of the transcripts corresponding to the participant. Additionally, we combine all participant speech segments corresponding to a single picture description for extracting acoustic features.

We extract 509 manually-engineered features from transcripts and associated audio files (see **Tables 4–6**). These features are identified as indicators of cognitive impairment in previous literature, and hence encode domain knowledge.

All the features are divided into three higher-level categories:

1. **Lexico-syntactic features (297):** Frequencies of various production rules from the constituency parsing tree of the transcripts (Chae and Nenkova, 2009), speech-graph based features (Mota et al., 2012), lexical norm-based features (e.g., average sentiment valence of all words in a transcript, average imageability of all words in a transcript; Warriner et al., 2013), features indicative of lexical richness. We also extract syntactic features (Ai and Lu, 2010) such as the proportion of various POS-tags, and similarity between consecutive utterances.
2. **Acoustic and temporal features (187):** Mel-frequency cepstral coefficients (MFCCs), fundamental frequency,

TABLE 4 | Summary of all lexico-syntactic features extracted.

Feature type	#Features	Brief Description
Syntactic complexity	36	L2 Analyzer features; utterance length, depth of syntactic parse tree
Production rules	104	Proportion of production type
Phrasal type ratios	13	Proportion, average length and rate of phrase types
Lexical norm-based	12	Average lexical norms across words for (e.g., imageability)
Lexical richness	6	Type-token ratios; brunet; Honor's statistic
Word category	5	Proportion of demonstratives, function words, Light verbs and inflected verbs, and propositions
Noun ratio	3	Ratios nouns:(nouns+verbs); nouns:verbs; pronouns:(nouns+pronouns)
Length measures	1	Average word length
Universal POS proportions	18	Proportions of Spacy universal POS tags
POS tag proportions	53	Proportions of Penn Treebank POS tags
Local coherence	15	Similarity between word2vec representations of utterances
Utterance distances	5	Fraction of pairs of utterances below a similarity threshold (0.5, 0.3, 0); avg/min distance
Speech-graph features	13	Representing words as nodes in a graph and computing density, number of loops, etc.
Utterance cohesion	1	Number of switches in verb tense across utterances divided by total number of utterances
Rate	2	Ratios—number of words: duration of audio; number of syllables: duration of speech,
Invalid words	1	Proportion of words not in the English dictionary
Sentiment norm-based	9	Average sentiment norms across all words, noun, and verbs

The number of features in each subtype is shown in the second column (titled "#Features").

TABLE 5 | Summary of all acoustic/temporal features extracted.

Feature type	#Features	Brief description
Pauses and fillers	9	Total and mean duration of pauses; long and short pause counts; pause to word ratio; fillers (um, uh); duration of pauses to word durations
Fundamental frequency	4	Avg/min/max/median fundamental frequency of audio
Duration-related	2	Duration of audio and spoken segment of audio
Zero-crossing rate	4	Avg/variance/skewness/kurtosis of zero-crossing rate
MFCC	168	Avg/variance/skewness/kurtosis of 42 MFCC coefficients

The number of features in each subtype is shown in the second column (titled "#Features").

statistics related to zero-crossing rate, as well as proportion of various pauses (for example, filled and unfilled pauses, ratio of a number of pauses to a number of words etc.; Davis and Maclagan, 2009).

TABLE 6 | Summary of all semantic features extracted.

Feature type	#Features	Brief description
Word frequency	10	Proportion of lemmatized words occurrences
Global coherence	15	Cosine distances between word2vec utterances and content units

The number of features in each subtype is shown in the second column (titled "#Features").

3. **Semantic features based on picture description content (25):** Proportions of various information content units used in the picture, identified as being relevant to memory impairment in prior literature (Croisile et al., 1996).

2.3. Experiments

2.3.1. AD vs. Non-AD Classification

2.3.1.1. Training Regimes

We benchmark the following training regimes for classification: classifying features extracted at transcript-level and a BERT model fine-tuned on transcripts.

Domain knowledge-based approach: We classify lexicosyntactic, semantic, and acoustic features extracted at transcript-level with four conventional ML models (SVM), neural network (NN), random forest (RF), naïve Bayes (NB)[1].

Hyperparameter tuning: All parameters in classification models were tuned to the best possible setting by searching within a grid of possible parameter values using 10-fold cross validation on the ADReSS challenge "train" set.

The random forest classifier fits 200 decision trees and considers $\sqrt{features}$ when looking for the best split. The minimum number of samples required to split an internal node is 2, and the minimum number of samples required to be at a leaf node is 2. Bootstrap samples are used when building trees. All other parameters are set to the default value.

The Gaussian Naive Bayes classifier is fit with balanced priors and variance smoothing coefficient set to $1e - 10$ and all other parameters default in each case.

The SVM is trained with a radial basis function kernel with kernel coefficient(γ) 0.001, and regularization parameter set to 100.

The NN used consists of two layers of 10 units each (note we varied both the number of units and number of layers while tuning for the optimal hyperparameter setting). The ReLU activation function is used at each hidden layer. The model is trained using Adam (Kingma and Ba, 2014) for 200 epochs and with a batch size of number of samples in train set in each fold. All other parameters are default.

We perform feature selection by choosing top-k number of features, based on ANOVA F-value between label/features. The number of features is jointly optimized with the classification model parameters.

[1] https://scikit-learn.org/stable/.

Transfer learning-based approach: In order to leverage the language information encoded by BERT (Devlin et al., 2019), we use pre-trained model weights to initialize our classification model. All our experiments are based on the *bert-base-uncased* variant (Devlin et al., 2019), which consists of 12 layers, each having a hidden size of 768 and 12 attention heads. Maximum input length is 512 tokens. Initial learning rate is set to $2e-5$, and Adam optimizer (Kingma and Ba, 2014) is used. Cross-entropy loss is used while fine-tuning for AD detection.

While the base BERT model is pre-trained with sentence pairs, our input to the model consists of speech transcripts with several transcribed utterances with start and separator special tokens from the BERT vocabulary at the beginning and end of each utterance respectively, following Liu and Lapata (2019). This is performed to ensure that utterance boundaries are easily encoded, since cross-utterance information such as coherence and utterance transitions is important for reliable AD detection (Fraser et al., 2016). An embedding, following Devlin et al. (2019), pooling information across all tokenized units in the transcript is extracted as the aggregate transcript representation from the BERT base for each transcript. This is then passed to the classification layer, and the combined model is fine-tuned on the AD detection task—all using an open-source PyTorch (Paszke et al., 2019) implementation of BERT-based text sequence classification models and tokenizers (Wolf et al., 2019). As noted by Devlin et al. (2019), this pooled embedding representation heavily depends on the fine-tuning task—in our case, AD detection at transcript level.

The transcript input to the classification model consists of several transcribed utterances with corresponding start and end tokens for each utterance, following (Liu and Lapata, 2019). The final hidden state corresponding to the first start (*[CLS]*) token in the transcript which summarizes the information across all tokens in the transcript using the self-attention mechanism in BERT is used as the aggregate representation, and passed to the classification layer (Devlin et al., 2019; Wolf et al., 2019). This model is then fine-tuned on training data.

Hyperparameter tuning: We optimize the number of epochs to 10 by varying it from 1 to 12 during CV. Adam optimizer (Kingma and Ba, 2014) and linear scheduling for the learning rate (Paszke et al., 2019) are used. Learning rate and other parameters are set based on prior work on fine-tuning BERT (Devlin et al., 2019; Wolf et al., 2019).

2.3.1.2. Evaluation

Cross-validation on ADReSS train set: We use two CV strategies in our work—leave-one-subject-out CV (LOSO CV) and 10-fold CV at transcript level. We report evaluation metrics with LOSO CV for all models except fine-tuned BERT for direct comparison to challenge baselines. Due to computational constraints of GPU memory, we are unable to perform LOSO CV for the BERT model. Hence, we perform 10-fold CV to compare feature-based classification models with fine-tuned BERT. Values of performance metrics for each model are averaged across three runs with different random seeds in all cases.

Predictions on ADReSS test set: We generate three predictions with different seeds from each hyperparameter-optimized classifier trained on the complete train set, and then produce a majority prediction to avoid overfitting. We report performance on the challenge test set, as obtained from the challenge organizers. We evaluate task performance primarily using accuracy scores, since all train/test sets are known to be balanced. We also report precision, recall, specificity, and F1 with respect to the positive class (AD).

2.3.2. MMSE Score Regression

2.3.2.1. Training regimes

Domain knowledge-based approach: For this task, we benchmark two kinds of regression models, linear, and ridge, using pre-engineered features as input. MMSE scores are always within the range of 0–30, and so predictions are clipped to a range between 0 and 30.

Hyperparameter tuning: Each model's performance is optimized using hyperparameters selected via grid-search LOSO CV. We perform feature selection by choosing top-k number of features, based on an F-Score computed from the correlation of each feature with MMSE score. The number of features is optimized for all models. For ridge regression, the number of features is jointly optimized with the coefficient for L2 regularization, α.

2.3.2.2. Evaluation

We report root mean squared error (RMSE) and mean absolute error (MAE) for the predictions produced by each of the models on the training set with LOSO CV. In addition, we include the RMSE for two models' predictions on the ADReSS test set. Hyperparameters for these models were selected based on performance in grid-search 10-fold cross validation on the training set, motivated by the thought that 10-fold CV better demonstrates how well a model will generalize to the test set.

3. RESULTS

3.1. AD vs. Non-AD Classification

In **Table 7**, the classification performance with all the models evaluated on the train set via 10-fold CV is displayed. We observe that BERT numerically outperforms all domain knowledge-based ML models with respect to all metrics, with an average accuracy of 81.8%. SVM is the best-performing domain knowledge-based model. However, accuracy of the fine-tuned BERT model is not significantly higher than that of the SVM classifier based on an Kruskal-Wallis H-test ($H = 0.4838, p > 0.05$). Note that we used a Kruskal-Wallis H-test here, and in performance-comparisons in sections below since we observe that accuracy is not normally distributed on varying the random seed while training/inference.

We also report the performance of all our classification models with LOSO CV (**Table 9**). Each of our classification models significantly outperform the challenge baseline, which is uses 34 simple language summary statistic measures (e.g., duration, total utterances, MLU, type-token ratio, percentages of nine parts of speech) on the CHAT transcripts by a large margin (+10% accuracy for the best performing model, $p = 0.036$ with Kruskal-Wallis H = 4.35 test). Feature selection results in accuracy increase of about 13% for the SVM classifier.

Performance results on the unseen, held out challenge test set are shown in **Table 8** and follow the trend of the cross-validated performance in terms of accuracy, with BERT outperforming the best feature-based classification model SVM with an accuracy of 83.33%, but not significantly so ($H = 2.4$, $p > 0.05$). The accuracy with a BERT-based classification model ranges between 85.14 and 81.25%.

3.2. MMSE Score Regression

Performance of regression models evaluated on both train and test sets is shown in **Table 9**. Ridge regression with 25 features selected attains the lowest RMSE of 4.56 (with a corresponding MAE of 3.50, or 11.67% error) during LOSO-CV on the training set. The results show that feature selection is impactful for performance and helps achieve a decrease of up to 1.5 RMSE points (and up to 0.86 of MAE) for a ridge regressor. Furthermore, a ridge regressor is able to achieve an RMSE of 4.56 on the ADReSS test set, a decrease of 0.64 from the baseline. We also experimented with different non-linear regression methods—however, given the small dataset size and the difficulty of the task, the linear regression models highlighted in **Table 9** performed the best.

4. DISCUSSION

4.1. Feature Differentiation Analysis

While we extracted a large number of linguistic and acoustic features to capture a wide range of linguistic and acoustic changes in speech associated with AD, based on a survey of prior literature (Yancheva et al., 2015; Fraser et al., 2016; Pou-Prom and Rudzicz, 2018; Zhu et al., 2019), we are also interested in identifying the *most differentiating* features between AD and non-AD speech. In order to study statistically significant differences in linguistic/acoustic phenomena, we perform independent t-tests between feature means for each class in the ADReSS training set, following the methodology followed by Eyre et al. (2020). 87 features are significantly different between the two groups at $p < 0.05$. Seventy-nine of these are text-based lexicosyntactic and semantic features, while eight are acoustic. These eight acoustic features include the number of long pauses, pause duration, and mean/skewness/variance-statistics of various MFCC coefficients. However, after Bonferroni correction for multiple testing, we identify that only 13 features are significantly different between AD and non-AD speech at $p < 9e - 5$, and none of these features are acoustic (**Table 10**). This implies that linguistic features are particularly differentiating between the AD/non-AD classes here, which explains why models trained only on linguistic features (i.e., BERT models) attain performance well above random chance.

The features that differentiate the AD and non-AD groups largely indicate semantic impairments in AD, reflected in the types of words used and the content of their picture descriptions. Importantly, many of the differentiating features replicate findings from Fraser et al. (2016), suggesting that despite the present dataset being more demographically balanced, many of the previous findings maintain. In addition, the differentiating features are consistent with other previous clinical literature

TABLE 7 | Ten-fold CV results averaged across three runs with different random seeds on the ADReSS train set.

Model	#Features	Accuracy	Precision	Recall	Specificity	F1
SVM	10	0.796	0.81	0.78	0.82	0.79
NN	10	0.762	0.77	0.75	0.77	0.76
RF	50	0.738	0.73	0.76	0.72	0.74
NB	80	0.750	0.76	0.74	0.76	0.75
BERT	–	**0.818**	**0.84**	**0.79**	**0.85**	**0.81**

Accuracy for BERT is higher, but not significantly so from SVM ($H = 0.4838, p > 0.05$ Kruskal-Wallis H-test). Bold indicates the best result.

TABLE 9 | LOSO-CV MMSE regression results on the ADReSS train and test sets.

Model	#Features	α	RMSE	MAE	RMSE
			Train set		Test set
Baseline (Luz et al., 2020)	–	–	4.38		5.20
LR	15	–	5.37	4.18	4.94
LR	20	–	4.94	3.72	–
Ridge	509	12	6.06	4.36	–
Ridge	35	12	4.87	3.79	**4.56**
Ridge	25	10	**4.56**	**3.50**	–

Bold indicates the best result.

TABLE 8 | AD detection results on unseen, held out ADReSS test set averaged over three runs with different random seeds.

Model	#Features	Accuracy	Precision	Recall	Specificity	F1	AUROC
Baseline (Luz et al., 2020)	–	0.7500	–	–	–	0.7800	–
SVM	10	0.8125	0.8000	0.8333	0.7917	0.8124	0.8125
NN	10	0.7708	0.7671	0.7778	0.7639	0.7708	0.7708
RF	50	0.7569	0.8033	0.6806	0.8333	0.7555	0.7500
NB	80	0.7292	0.7895	0.6250	0.8333	0.7262	0.7292
BERT	–	**0.8332**	**0.8389**	**0.8333**	**0.8333**	**0.8327**	**0.8333**

Bold indicates the best result.

TABLE 10 | Feature differentiation analysis results for the most important features, based on ADReSS train set.

Feature	Feature type	μ_{AD}	μ_{non-AD}	Correlation	Weight
Average cosine distance between utterances	Semantic	0.91	0.94	–	–
Fraction of pairs of utterances below a similarity threshold (0.5)	Semantic	0.03	0.01	–	–
Cosine distance between word2vec utterances and content units	Semantic	0.46	0.38	−0.54*	−1.01
Distinct content units mentioned: total content units	Semantic	0.27	0.45	0.63*	1.78
Distinct action content units mentioned: total content units	Semantic	0.15	0.30	0.49*	1.04
Distinct object content units mentioned: total content units	Semantic	0.28	0.47	0.59*	1.72
Cosine distance between GloVe utterances and content units	Semantic	–	–	−0.42*	−0.03
Average word length (in letters)	Lexico-syntactic	3.57	3.78	0.45*	1.07
Proportion of pronouns	Lexico-syntactic	0.09	0.06	–	–
Ratio (pronouns):(pronouns+nouns)	Lexico-syntactic	0.35	0.23	–	–
Proportion of personal pronouns	Lexico-syntactic	0.09	0.06	–	–
Proportion of adverbs	Lexico-syntactic	0.06	0.04	−0.41*	−0.41
Proportion of adverbial phrases amongst all rules	Lexico-syntactic	0.02	0.01	−0.37	−0.74
Proportion of non-dictionary words	Lexico-syntactic	0.11	0.08	–	–
Proportion of gerund verbs	Lexico-syntactic	–	–	0.37	1.08
Proportion of words in adverb category	Lexico-syntactic	–	–	−0.4*	−0.49

*μ_{AD} and μ_{non-AD} show the means of the 13 significantly different features at $p < 9e\text{-}5$ (after Bonferroni correction) for the AD and non-AD group, respectively. We also show Spearman correlation between MMSE score and features, and regression weights of the features associated with the five greatest and five lowest regression weights from our regression experiments. *Next to correlation indicates significance at $p < 9e\text{-}5$.*

documenting decreased specificity and information content in AD. For example, the features relating to the content units in the picture and the cosine similarity between utterances and picture content units show that the picture descriptions produced in AD have fewer relevant content words and that the words used are less semantically related to the themes of the picture. Lower average cosine distance in AD signifies more repetition in speech. These findings are consistent with previous studies reporting reduced information content and coherence in AD (Croisile et al., 1996; Snowdon et al., 1996; Dijkstra et al., 2004; Forbes-McKay and Venneri, 2005; Riley et al., 2005; Le et al., 2011; Ahmed et al., 2013; Boschi et al., 2017). Other differentiating features related to the use of shorter words, and increased use of pronouns, adverbs, and words not found in the dictionary. These features may all reflect the use of less specific and simpler language, and replicate previous findings of decreased specificity of language in AD (Le et al., 2011; Ahmed et al., 2013; Szatloczki et al., 2015; Fraser et al., 2016). Interestingly, while Fraser et al. (2016) found differences in acoustic features, none of those findings survived Bonferroni correction in the present study, which may indicate that this age/sex-balanced dataset reduced the acoustic differences between groups.

In order to visualize the class-separability of the feature-based representations, we visualize (t-SNE) t-Distributed Stochastic Neighbor Embedding (Maaten and Hinton, 2008) plots in **Figure 1**. t-SNE is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It maps multi-dimensional data to two or more dimensions suitable for human observation. We observe strong class-separation between the two classes, indicating that a non-linear model would be capable of good AD detection performance with these representations.

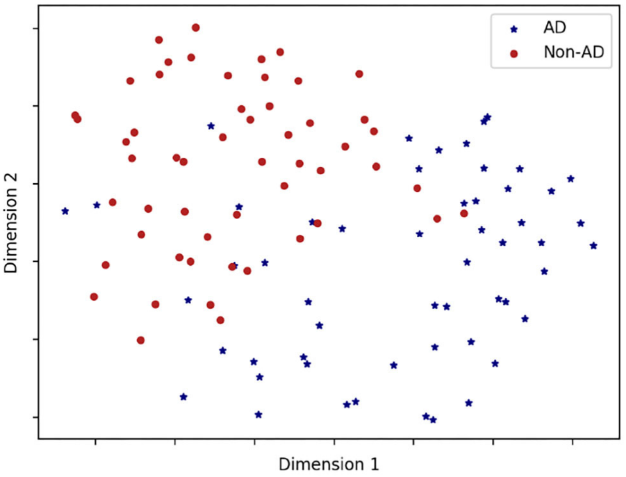


FIGURE 1 | A t-SNE plot showing class separation. Note we only use the 13 features significantly different between classes (see **Table 10**) in feature representation for this plot.

4.2. Interpreting Attention Patterns in BERT-Based Models

We look at multi-scale attention visualizations of BERT fine-tuned for the AD detection task, using the BertViz library (Vig, 2019) (**Figure 2**). Self-attention is an important component of BERT-based models, and looking at attention patterns can help us interpret model decisions. We used the BERT-base model which consists of 12 layers, and 12 attention heads in each layer. We visualize, for both AD and healthy speech transcripts, the attention weights for the final “[CLS]” token,

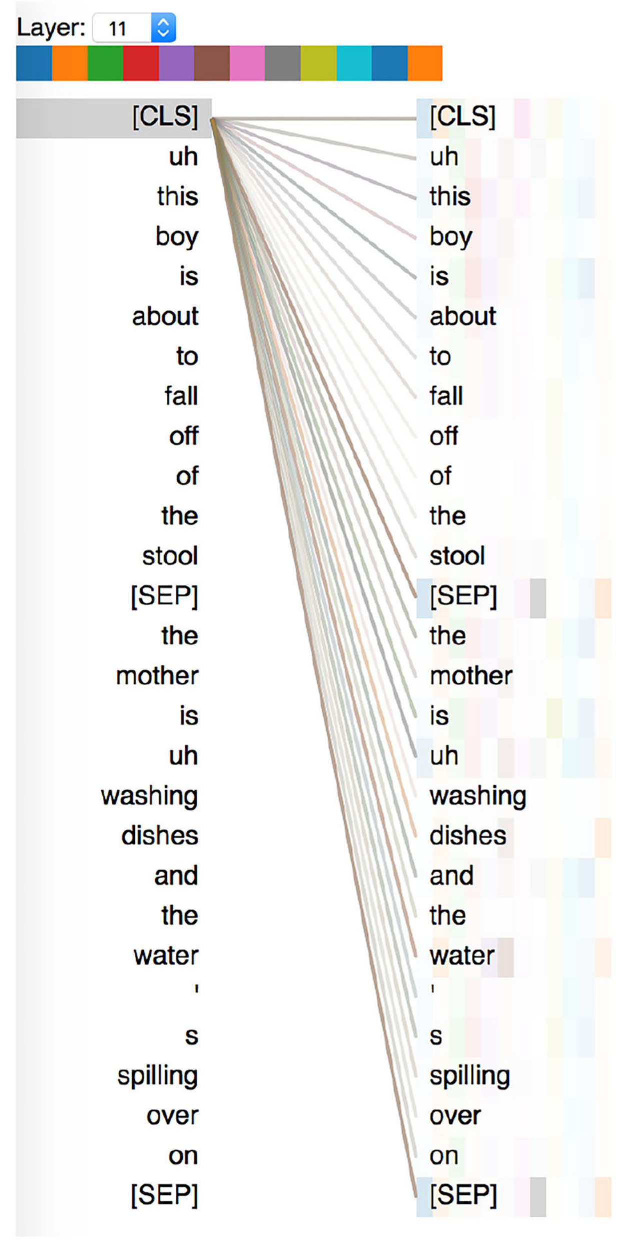


FIGURE 2 | An attention visualization plot showing attention contributions of embeddings corresponding to each word to the "pooled" representation. This example is a sub-sample (first two utterances) of a speech transcript from a healthy person.

whose representation is passed to the fully-connected layer for classification. On analyzing the attention weights attributed to words in both healthy and AD transcripts, we find that:

1. attention weights are often attributed to a few important "information content units." which have been identified to be important speech indicators of AD in prior work (Fraser et al., 2016) such as "water," "boy," etc.
2. attention weights are also sometimes attributed to pauses and fillers, such as "uh" and "um."

TABLE 11 | LOSO-CV results averaged across three runs with different random seeds on the ADReSS train set.

Model	#Features	Accuracy	Precision	Recall	Specificity	F1
Baseline (Luz et al., 2020)	–	0.768	0.77	0.76	–	0.77
SVM	509	0.741	0.75	0.72	0.76	0.74
SVM	10	**0.870**	**0.90**	**0.83**	**0.91**	**0.87**
NN	10	0.836	0.86	0.81	0.86	0.83
RF	50	0.778	0.79	0.77	0.79	0.78
NB	80	0.787	0.80	0.76	0.82	0.78

Accuracy for SVM is significantly higher than NN ($H = 4.50, p = 0.034$ Kruskal-Wallis H-test). Bold indicates the best result.

3. attention weights are also attributed to the sentence separator tokens, and we think this approximates to roughly counting the number of utterances in the transcript.

Hence, as seen in sections 4.1 and 4.2, we observe that for both the feature-based classification models and BERT-based models, information units and fillers such as "uh" and "um" seem to be important predictors, similar to findings observed by Yuan et al. (2020).

4.3. Analysing AD Detection Performance Differences

We observe that both feature-based and BERT-based classification models are significantly better than the linguistic baseline, showing the importance of an extensive amount of linguistic features for detecting AD-related differences. When compared on this well-matched dataset, BERT tended to have higher performance, but the difference was not significant. Based on feature differentiation analysis, we hypothesize that good performance with a text-focused BERT model on this speech classification task is due to the strong utility of linguistic features on this dataset. BERT captures a wide range of linguistic phenomena due to its training methodology, potentially encapsulating most of the important lexico-syntactic and semantic features. It is thus able to use information present in the lexicon, syntax, and semantics of the transcribed speech after fine-tuning (Jawahar et al., 2019).

We also see a trend of better performance when increasing the number of folds (see SVM in **Tables 7**, **11**) in cross-validation. We postulate that this is due to the small size of the dataset, and hence differences in training set size in each fold ($N_{train} = 107$ with LOSO, $N_{train} = 98$ with 10-fold CV). Note that, in this dataset, both feature-based and BERT-based classification methods rely on linguistic features to achieve better classification than baseline. This implies that the linguistic features from speech transcripts are quite informative for the AD detection task. Hence, an interesting direction of future research is expanding our current set of features to incorporate more discourse-related features (which could be getting captured to some degree in fine-tuned BERT models).

4.4. Regression Weights for MMSE Prediction

To assess the relative importance of individual input features for MMSE prediction, we report features with the five highest and five lowest regression weights reflecting the five strongest positive and negative relationships with MMSE scores (**Table 10**). Each presented value is the average weight assigned to that feature across each of the LOSO CV folds. We also present the correlation with MMSE score coefficients for those 10 features, as well as their significance, in **Table 10**. We observe that for each of these highly weighted features, a positive or negative correlation coefficient is accompanied by a positive or negative regression weight, respectively. This demonstrates that these 10 features are so distinguishing that, even in the presence of other regressors, their relationship with MMSE score remains the same. We also note that all 10 of these are linguistic features, further demonstrating that linguistic information is particularly distinguishing when it comes to predicting the severity of a patient's AD. Notably, seven of the ten features were among those that differentiated between AD and non-AD groups, demonstrating that there is high overlap between the features relevant to group differentiation and MMSE score prediction. These features included those relating to the information content and the coherence of picture descriptions, reflected by content unit and cosine distance features. Word length and use of adverbs were also relevant to MMSE prediction, with longer words and fewer adverbs correlating with higher MMSE scores. The use of gerund verbs was found to have a high regression weight for MMSE prediction and positively correlated with MMSE scores, despite not being significantly different between AD and non-AD groups after Bonferroni correction. Reduced use of inflected verbs has been found in some previous research (Ahmed et al., 2013; Fraser et al., 2016), and is thought to reflect an grammatic impairment.

5. CONCLUSIONS

In this paper, we rigorously compare two widely used approaches—linguistic and acoustic feature engineering based on domain knowledge, and text-only transfer learning using fine-tuned BERT classification model. Our results show that pre-trained models that are fine-tuned for the AD classification task are capable of performing well on AD detection, achieving comparable, or even slightly improved performance compared to hand-crafted feature engineering. We observe that linguistic features are capable of attaining predictive performance well above chance on this acoustically and demographically balanced speech dataset, and posit this to be the reason why a text-only approach with BERT numerically outperforms a multi-modal feature-engineering based approach. The present findings highlight the importance of measuring the linguistic, and especially semantic content of speech, in addition to acoustic analyses. In future work, it would be interesting to study methods that combine feature-based and pre-trained neural LM-based prediction models to optimize AD detection from speech—this could potentially help harness complementary benefits of both approaches. It is interesting to note that the winners of the ADReSS challenge also used a pre-trained language model, augmented with additional information about speech disfluencies (Yuan et al., 2020), which outperforms our best model by 6% in accuracy and F1-score, further indicating the degree of promise in such an approach. These results build on previous work to demonstrate how automated speech analysis can be used to help characterize AD. Speech samples can be collected quickly and non-invasively, and as demonstrated in the present results, yield measures relating to the presence and severity of AD.

Further work will build on these results to develop improved tools for disease screening and monitoring in AD, improving the efficiency of clinical research and treatment. In the future, we will experiment with different neural models such as XLNet (Yang et al., 2019), and with different tokenization and encoding strategies for transcript representations. A direction for future work is developing ML models that combine representations from BERT and hand-crafted features (Yu et al., 2015). Such feature-fusion approaches could potentially boost performance on the cognitive impairment detection task.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://dementia.talkbank.org/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by DementiaBank consortium. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to writing and edits. Methods and analyses were performed by AB, JN, and BE.

FUNDING

FR was supported by a CIFAR Chair in AI.

ACKNOWLEDGMENTS

The results shown in this manuscript were also presented at INTERSPEECH, 2020 as a part of the ADReSS challenge track (Balagopalan et al., 2020).

REFERENCES

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* 136, 3727–3737. doi: 10.1093/brain/awt269

Ai, H., and Lu, X. (2010). "A web-based system for automatic measurement of lexical complexity," in *27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10)* (Amherst, MA), 8–12.

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. *Proc. Interspeech* 2020, 2167–2171. doi: 10.21437/Interspeech.2020-2557

Balagopalan, A., Novikova, J., Rudzicz, F., and Ghassemi, M. (2018). The effect of heterogeneous data for Alzheimer's disease detection from speech. *arXiv preprint arXiv:1811.12254*.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol*. 51, 585–594. doi: 10.1001/archneur.1994.00540180063015

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., and Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: a review. *Front. Psychol*. 8:269. doi: 10.3389/fpsyg.2017.00269

Chae, J., and Nenkova, A. (2009). "Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text," in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (Athens), 139–147. doi: 10.3115/1609067.1609082

Chen, L., Dodge, H. H., and Asgari, M. (2020). "Topic-based measures of conversation for detecting mild cognitive impairment," in *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations* (Virtual), 63–67.

Cockrell, J. R., and Folstein, M. F. (2002). Mini-mental state examination. *Princ. Pract. Geriatr. Psychiatry*, 140–141. doi: 10.1002/0470846410.ch27(ii)

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., et al. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain Lang*. 53, 1–19. doi: 10.1006/brln.1996.0033

Davis, B. H., and MacLagan, M. (2009). Examining pauses in Alzheimer's discourse. *Am. J. Alzheimer's Dis. Other Dement*. 24, 141–154. doi: 10.1177/1533317508328138

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J. Alzheimer's Dis*. 78, 1547–1574. doi: 10.3233/JAD-200888

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Volume 1 (Long and Short Papers)* (Minneapolis, MN), 4171–4186.

Dijkstra, K., Bourgeois, M. S., Allen, R. S., and Burgio, L. D. (2004). Conversational coherence: discourse analysis of older adults with and without dementia. *J. Neurolinguist*. 17, 263–283. doi: 10.1016/S0911-6044(03)00048-4

Eyre, B., Balagopalan, A., and Novikova, J. (2020). "Fantastic features and where to find them: detecting cognitive impairment with a subsequence classification guided approach," in *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020)* (Virtual), 193–199. doi: 10.18653/v1/2020.wnut-1.25

Forbes-McKay, K. E., and Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurol. Sci*. 26, 243–254. doi: 10.1007/s10072-005-0467-9

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimer's Dis*. 49, 407–422. doi: 10.3233/JAD-150520

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *BDAE-3: Boston Diagnostic Aphasia Examination, 3rd Edn*. Philadelphia, PA: Lippincott Williams & Wilkins.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., and Hoffmann, I. (2019). Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Comput. Speech Lang*. 53, 181–197. doi: 10.1016/j.csl.2018.07.007

Jammeh, E. A., Camille, B. C., Stephen, W. P., Escudero, J., Anastasiou, A., Zhao, P., et al. (2018). Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. *BJGP Open* 2:bjgpopen18X101589. doi: 10.3399/bjgpopen18X101589

Jawahar, G., Sagot, B., and Seddah, D. (2019). "What does bert learn about the structure of language?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence), 3651–3657. doi: 10.18653/v1/P19-1356

Johnson, A. E., Pollard, T. J., and Naumann, T. (2018). Generalizability of predictive models for intensive care unit patients. *arXiv preprint arXiv:1812.02275*.

Karlekar, S., Niu, T., and Bansal, M. (2018). "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Volume 2 (Short Papers)* (New Orleans, LA), 701–707. doi: 10.18653/v1/N18-2110

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's Dement. Diagn. Assess. Dis. Monit*. 1, 112–124. doi: 10.1016/j.dadm.2014.11.012

Le, X., Lancashire, I., Hirst, G., and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Liter. Linguist. Comput*. 26, 435–461. doi: 10.1093/llc/fqr013

Liu, Y., and Lapata, M. (2019). "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong), 3721–3731. doi: 10.18653/v1/D19-1387

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the address challenge. *arXiv:2004.06833*. doi: 10.21437/Interspeech.2020-2571

Maaten, L. v. d., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res*. 9, 2579–2605.

MacWhinney, B. (2000). The CHILDES project: tools for analyzing talk: Volume I: Transcription format and programs, Volume II: the database. *Comput. Linguist*. 26:657. doi: 10.1162/coli.2000.26.4.657

Mota, N. B., Vasconcelos, N. A., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., et al. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS ONE* 7:e34928. doi: 10.1371/journal.pone.0034928

Noorian, Z., Pou-Prom, C., and Rudzicz, F. (2017). On the importance of normative data in speech-based assessment. *arXiv preprint arXiv:1712.00069*.

Orimaye, S. O., Tai, K. Y., Wong, J. S.-M., and Wong, C. P. (2015). Learning linguistic biomarkers for predicting mild cognitive impairment using compound skip-grams. *arXiv preprint arXiv:1511.02436*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* (Vancouver, CA), 8024–8035.

Petti, U., Baker, S., and Korhonen, A. (2020). A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J. Am. Med. Inform. Assoc*. 27, 1784–1797. doi: 10.1093/jamia/ocaa174

Pou-Prom, C., and Rudzicz, F. (2018). "Learning multiview embeddings for assessing dementia," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels), 2812–2817. doi: 10.18653/v1/D18-1304

Prabhakaran, G., and Bakshi, R. (2018). Analysis of structure and cost in a longitudinal study of Alzheimer's disease. *J. Health Care Fin*. 8:411. doi: 10.4172/2161-0460.1000411

Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M., and Karagiannidou, M. (2016). *World Alzheimer Report 2016: Improving Healthcare for People Living With Dementia: Coverage, Quality and Costs Now and in the Future*. Alzheimer's Disease International.

Pulido, M. L. B., Hernández, J. B. A., Ballester, M. Á. F., González, C. M. T., Mekyska, J., and Smékal, Z. (2020). Alzheimer's disease and automatic speech analysis: a review. *Expert Syst. Appl*. 150:113213. doi: 10.1016/j.eswa.2020.113213

Riley, K. P., Snowdon, D. A., Desrosiers, M. F., and Markesbery, W. R. (2005). Early life linguistic ability, late life cognitive function, and neuropathology: findings from the nun study. *Neurobiol. Aging* 26, 341–347. doi: 10.1016/j.neurobiolaging.2004.06.019

Rogers, A., Kovaleva, O., and Rumshisky, A. (2021). A primer in bertology: what we know about how bert works. *Trans. Assoc. Comput. Linguist*. 8, 842–866. doi: 10.1162/tacl_a_00349

Slegers, A., Filiou, R.-P., Montembeault, M., and Brambati, S. M. (2018). Connected speech features from picture description in Alzheimer's disease: a systematic review. *J. Alzheimer's Dis*. 65, 519–542. doi: 10.3233/JAD-170881

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the nun study. *JAMA* 275, 528–532. doi: 10.1001/jama.1996.03530310034029

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Front. Aging Neurosci*. 7:195. doi: 10.3389/fnagi.2015.00195

Vig, J. (2019). A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*. doi: 10.18653/v1/P19-3007

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* 45, 1191–1207. doi: 10.3758/s13428-012-0314-x

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2019). Huggingface's transformers: state-of-the-art natural language processing. *ArXiv abs/1910.03771*. doi: 10.18653/v1/2020.emnlp-demos.6

Yancheva, M., Fraser, K. C., and Rudzicz, F. (2015). "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies* (Dresden), 134–139. doi: 10.18653/v1/W15-5123

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). "Xlnet: generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems* (Vancouver, CA), 5753–5763.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag*. 13, 55–75. doi: 10.1109/MCI.2018.2840738

Yu, M., Gormley, M. R., and Dredze, M. (2015). "Combining word embeddings and feature embeddings for fine-grained relation extraction," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics* (Denver, CO), 1374–1379. doi: 10.3115/v1/N15-1155

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. *Proc. Interspeech* 2020, 2162–2166. doi: 10.21437/Interspeech.2020-2516

Zhu, Z., Novikova, J., and Rudzicz, F. (2018). Semi-supervised classification by reaching consensus among modalities. *arXiv preprint arXiv:1805.09366*.

Zhu, Z., Novikova, J., and Rudzicz, F. (2019). "Detecting cognitive impairments by agreeing on interpretations of linguistic features," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Minneapolis, MN), 1431–1441. doi: 10.18653/v1/N19-1146

Conflict of Interest: Authors AB, BE, JR and JN were employed by company Winterlight Labs Inc.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 29 April 2021
doi: 10.3389/fcomp.2021.649508

Analysis and Classification of Word Co-Occurrence Networks From Alzheimer's Patients and Controls

Tristan Millington * and Saturnino Luz*

Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, United Kingdom

In this paper we construct word co-occurrence networks from transcript data of controls and patients with potential Alzheimer's disease using the ADReSS challenge dataset of spontaneous speech. We examine measures of the structure of these networks for significant differences, finding that networks from Alzheimer's patients have a lower heterogeneity and centralization, but a higher edge density. We then use these measures, a network embedding method and some measures from the word frequency distribution to classify the transcripts into control or Alzheimer's, and to estimate the cognitive test score of a participant based on the transcript. We find it is possible to distinguish between the AD and control networks on structure alone, achieving 66.7% accuracy on the test set, and to predict cognitive scores with a root mean squared error of 5.675. Using the network measures is more successful than using the network embedding method. However, if the networks are shuffled we find relatively few of the measures are different, indicating that word frequency drives many of the network properties. This observation is borne out by the classification experiments, where word frequency measures perform similarly to the network measures.

Keywords: machine learning, natural language processing, Alzheimer's disease, network analysis, network embedding, graph measures

OPEN ACCESS

Edited by:
Anna Esposito,
University of Campania Luigi Vanvitelli,
Italy

Reviewed by:
Carmen Klaussner,
MTI Technology, Vietnam
Rytis Maskeliunas,
Kaunas University of Technology,
Lithuania

****Correspondence:***
Tristan Millington
tristan.millington@ed.ac.uk

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *04 January 2021*
Accepted: *13 April 2021*
Published: *29 April 2021*

Citation:
Millington T and Luz S (2021) Analysis and Classification of Word Co-Occurrence Networks From Alzheimer's Patients and Controls.
Front. Comput. Sci. 3:649508.
doi: 10.3389/fcomp.2021.649508

1 INTRODUCTION

As populations continue to age, the development of automated methods to help reduce the amount of in person care required is becoming an important research topic. Dementia is a particular issue, where the cognitive function of a person declines as they age, with symptoms including memory loss, motor problems, deterioration of visuospatial function, language impairment and emotional distress. These issues tend to reduce the ability of a person to care for themselves, placing an added burden on their carers and/or relatives. Early diagnosis of dementia is desirable as it is amenable to treatment, and this can help the patient live a longer, more independent life. Dementia shows various linguistic effects, with patients tending to produce sentences with less information, less syntactic complexity (Pakhomov et al., 2011), fewer unique words and more meaningless sentences (Fraser et al., 2016). These effects can be used for non-invasive diagnosis and analysis of dementia, and so in this paper we look at using text classification methods to this end.

The common approach in text classification is to use a bag of words model. This assumes that word order does not matter, and either counts the number of occurrences of each word in a document, or uses some information based measures such as term-frequency inverse document frequency (TF-IDF). Various authors have taken this approach, and demonstrated good results on classifying participants as AD or controls (Orimaye et al., 2017; Wankerl et al., 2017). However, word order does in fact matter, and we can try to capture this using graph based methods. The approach used here is to construct a graph where the words in the document are nodes, and if two words co-occur within a certain window (a set of words occurring around a given word) an edge is drawn between them. Furthermore, these co-occurrence networks are an approximation of syntactic networks, as most syntactic relationships occur between words that are close together (Cancho and Solé, 2001). Various syntactic measures have been previously used to distinguish between AD patients and controls (Pakhomov et al., 2011; Fraser et al., 2016), and we hypothesize that these co-occurrence networks can capture these syntactic relations without the use of a syntactic parser.

Therefore, in this paper we investigate the properties of word co-occurrence networks using a variety of co-occurrence windows from transcripts of controls and patients diagnosed with potential Alzheimer's disease (AD) on a picture description task. We analyze these networks for potential differences between the controls and AD patients using various network measures, and then look at classifying the networks using a set of network measures, a graph embedding method and a baseline method using word frequency statistics. Each transcript is also annotated with the mini-mental state examination (MMSE) result. This is a test of cognitive function, and can be used to help diagnose dementia. The test scores ranges from 0–30, and a score below 24 is usually taken to indicate cognitive impairment. We are also interested in predicting the value of the MMSE score from a transcript by using this co-occurrence network model and the graph measures/network embedding method. To the best of our knowledge, such an approach has not been taken before. We use the terms graph and network interchangeably in this paper, and we emphasize these networks we refer to are different to neural networks.

2 RELATED WORK

The structure of word co-occurrence networks has been studied by many authors, along with which parameters can be used for classification. For instance Liu and Cong (2013) study the use of various network measures for distinguishing between the same text written in a set of different languages. Focusing mostly on Slavic languages (although they do also use English) they use hierarchical clustering to show which languages are more similar. By trying many different combinations of the network measures, they discover that it is possible to show the Slavic languages are more similar to each other than they are to English or Chinese, and inside the Slavic group the languages that are generally regarded as more similar (e.g. Belorussian and Russian) are more closely clustered than less similar languages (e.g. Russian and Slovakian). Other authors have applied similar methods for author attribution (Antiqueira et al., 2007; Mehri et al., 2012; Akimushkin et al., 2017), distinguishing between automatically generated and human written text (Amancio et al., 2008; Amancio, 2015) and for keyphrase extraction (Mihalcea and Tarau, 2004; Bougouin et al., 2013; Florescu and Caragea, 2017). A detailed review of the literature so far on the construction and applications of word co-occurrence networks is provided by Cong and Liu (2014).

Graphs can also be used to augment n-gram classification. For instance, we can gain the centrality of a term from a graph, which can then be used as input into a text classification algorithm (Hassan et al., 2007), and this has been shown to improve classification accuracy compared to just using n-grams. Alternatively we can use the structure of the graph as input into the classification algorithm. This has the advantage of ensuring that new documents can have unknown words, which is advantageous if the system must be deployed for a period of time, as it is unlikely that every word that could ever be encountered is present in the training set. Rousseau et al. (2015) use the subgraph mining method gSpan (Yan and Han, 2002) to mine frequent subgraphs from a set of graphs extracted from text documents. The presence of these subgraphs is then used as input into the classification method. The disadvantage of this method is that it is computationally expensive to mine for all possible subgraphs of non-trivial size.

A similar approach to the one we take in this paper is proposed by Santos et al. (2017). In their paper the authors apply a word co-occurrence network model to the DementiaBank and a Portuguese dataset. However, unlike us they enrich their model using word embeddings to produce weighted edges between words that do not co-occur. They calculate node level statistics for each graph and use this as input into a classification procedure. With their enriched networks they achieve an increase in classification accuracy, achieving 62% on the DementiaBank dataset.

There have been many more approaches taken to identify Alzheimer's using machine learning techniques and linguistic features. One of the first examples in the literature is the analysis of the books of an author who was diagnosed with Alzheimer's by Garrard et al. (2005). A combination of lexical, syntactic and vocabulary based features is used to compare the books. This is further extended by Pakhomov et al. (2011). Using the Stanford parser, the authors take three measures of syntactic complexity, Yngve depth, Frazier depth and the length between grammatical dependencies. There is a clear decline over time in the syntactic complexity of the authors writing, particularly with the books at the end of her career.

Many authors have used the DementiaBank corpus for their studies. For instance, Fraser et al. (2016) apply machine learning methods to the DementiaBank corpus, using both transcripts and speech data. Firstly they used logistic regression to evaluate the contribute of each feature to successful classification of a participant as having Alzheimer's or being a control. Ranking the features using Pearson correlation, they firstly investigate how including more features affects the classification accuracy. The maximum classification accuracy is achieved when the 35 most correlated features are used (at 81.92%), and beyond this it tends to remain roughly constant until 50 features are reached (dropping slightly to 78.72%, after which the classification accuracy decreases significantly. Of relevance to this paper, they find that AD patients produce more pronouns, fewer nouns, have a smaller vocabulary and repeat themselves more.

Interestingly though, measures such as the depth of the parse tree do not seem significantly correlated with an AD diagnosis.

Orimaye et al. (2017) further explore which features can be used to distinguish between AD and controls from transcripts. A combination of n-grams, lexical and syntactic features to this end. Since this is a large feature set, they perform some univariate screening using t-tests to remove variables which have little ability to distinguish between classes. Of particular interest to us, they find that the number of repetitions, reduced sentences, predicates and mean length of utterances are different between the classes, but that many other syntactic measures, such as the dependency distance, are not. They then select the top 1,000 features for input into the SVM classifier. Comparing the syntactic and lexical features only, to the n-gram only, to the combination of the two, they find that the combination performs the best, with an AUC of 0.93, compared to 0.80 for the lexical-syntactic features and 0.91 for the n-gram features.

Authors have also disregarded syntactic features and used only n-grams for classification. For instance, Garrard et al. (2014) use n-grams to distinguish between AD patients and controls on a picture description task. They find that the transcripts from the controls contain more content words (e.g. picnic, blanket) while the AD patients tend to produce more generic terms (e.g. something, thing). They use only a small subset of the total word set to classify, as there are a large number of possible n-grams. This indicates that a small number of features can be used to distinguish between AD patients and controls.

Larger n-grams can also be used. Orimaye et al. (2018) use a deep neural network and large n-grams ($n > 3$) to classify the participants into control or AD. Since the occurrence matrix of these n-grams will be very sparse, they firstly reduce the dimensionality using SVD (selecting 19 features in the end), before inputting this smaller matrix into the neural network. Experimenting with a variety of n-gram sizes, they find using 4 g achieves the lowest error (11.1%) on the test set in their deep neural networks. It is also possible to use the distribution of n-grams to differentiate between AD and controls. Wankerl et al. (2017) create probability distributions of the trigrams in transcripts from the cookie detection task from DementiaBank. The perplexity of a new sample is used to classify it as AD or control.

While transcripts are convenient to analyze, transcription can be challenging, either noisy if done automatically or slow and expensive if done by humans. Using purely audio is attractive if we wish to apply these methods on non curated datasets. Haider et al. (2019) study the same corpus, but instead focus their efforts on purely acoustic features. Here they use a fusion of acoustic feature sets (namely emobase, ComParE, eGeMaps and MRCG) on the DementiaBank dataset, achieving a maximum accuracy of 78.8%. A challenge with many of these approaches is that they are dependent on language and context. To solve these issues, Luz et al. (2018) instead propose to extract vocalization graphs from patient dialogue. Using features from these vocalization graphs, they achieve a classification accuracy of 86.5%, though on a different dataset.

Aside from speech data, other approaches have included the use of smart home data (Alberdi et al., 2018). This particular example involves using activity recognition to establish routines, and then these routines can be compared between healthy participants and those with AD. If the reader is curious for more details, comprehensive reviews on the topic of Alzheimer's detection are provided by de la Fuente Garcia et al. (2020) and Slegers et al. (2018).

3 SOFTWARE AND DATA

Our dataset is made up of transcripts from the DementiaBank corpus. The DementiaBank corpus is a set of recordings of cognitive tests, which forms part of the larger TalkBank project (MacWhinney, 2019). The subset of DementiaBank used in this study encompasses recordings and their corresponding transcriptions, where patients with Alzheimer's and controls describe a picture known as the "Cookie Theft" scene, taken from the Boston Diagnostic Aphasia Examination (Becker et al. (1994)). This dataset is known as the Pitt corpus. Participants were required to:

- be above 44 years of age,
- have at least 7 years of education,
- have no history of nervous system disorders,
- not be taking neuroleptic medication,
- have an MMSE score of above 10.
- be able to give informed consent, and
- have a caregiver or relative to act as an informant if they had dementia.

To avoid possible biases due to age and gender which might have affected some of the above mentioned machine learning studies (de la Fuente Garcia et al., 2020), we use the ADReSS challenge dataset (Luz et al., 2020). The age and gender distributions of participants in DementiaBank's Pitt Corpus tend to reflect the fact that age and gender are major risk factors in AD. Therefore, AD participants will tend to be older and more likely to be female than control participants. The ADReSS dataset removes this source of bias as it consists of a subset of the Pitt corpus, sampled so as to be balanced with respect to gender and age. This dataset is divided into two halves, a training set and a test set. The training set contains 108 transcripts, evenly split between the AD and controls. The test set contains 48 transcripts, again evenly balanced between AD and controls. We perform our analysis on the training set, and keep the test set as an unseen dataset for evaluating the classifiers. The source code for the experiments described in this paper is available at our Gitlab repository.[1] Instructions on how to acquire the dataset are available at the ADReSS website.[2]

The networks are built using word co-occurrence windows of 2, 3 and 5, and are weighted and undirected. The weight on an edge is the number of times the two words occur together within a window in the same sentence. We remove any characters that are not in the Latin alphabet (i.e. numbers and punctuation are removed), but do not perform any stop word removal or

[1] https://git.ecdf.ed.ac.uk/tmilling/analysis-and-classification-of-word-co-occurrence-networks

[2] https://edin.ac/375QRNI

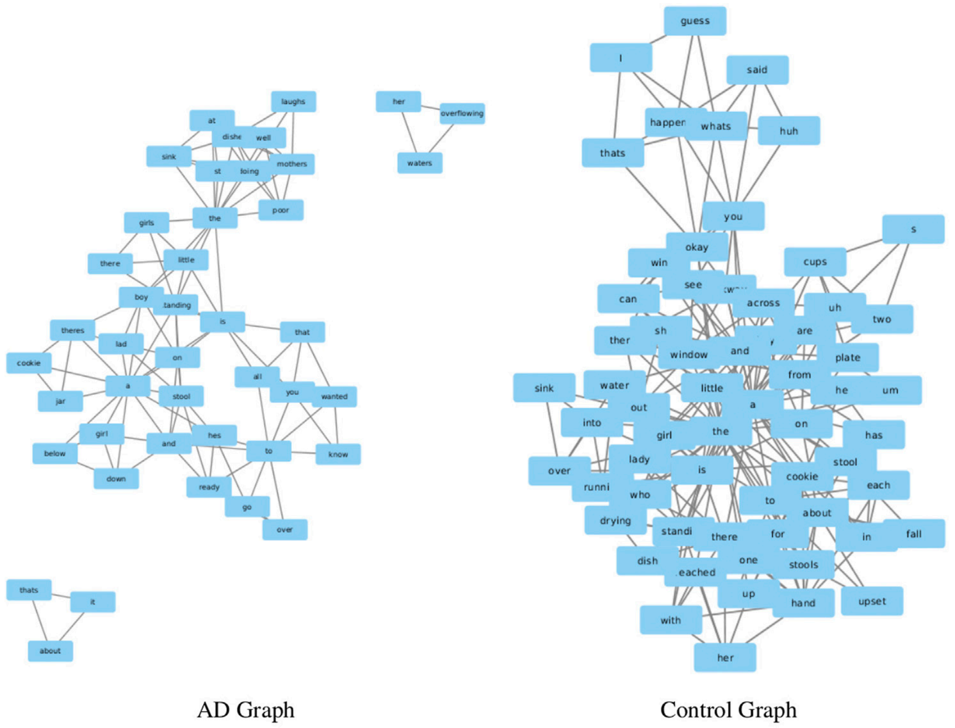


FIGURE 1 | Example graphs.

lemmatization. We do not remove stop words as we hope that the networks capture differences in their usage between the AD patients and controls. Pauses and other "non word" utterances are retained.

We make use of Python, NumPy and SciPy (Oliphant, 2006) for general scripting, pandas (McKinney, 2010) for handing the data, matplotlib (Hunter, 2007) for plotting, Networkx (Hagberg et al., 2008) for the network analysis, Cytoscape (Shannon et al., 2003) for the graph visualization, powerlaw (Alstott et al., 2014) for fitting power laws to the degree distributions, scikit-learn (Pedregosa et al., 2011) for implementation of the classifiers, NLTK (Loper and Bird, 2002) for some of the natural language processing, PyLangAcq (Lee et al., 2016) for parsing the transcriptions and Karate Club (Rozemberczki et al., 2020) for the graph embeddings.

4 NETWORK ANALYSIS

4.1 Method

To start with, we show example networks from the control and AD patients in **Figure 1**. Next we look at the values of various network measures for the co-occurrence networks constructed from the patients and controls. We only use the training set for this analysis. We focus on similar measures to previous work (Liu and Cong (2013)), in this case choosing.

- Number of nodes (N)
- Number of edges (E)
- Edge density (ED)
- Fraction of self links (SL)
- Average Clustering Coefficient ($\langle CC\rangle$)
- Diameter (D)
- Heterogeneity (how similar the nodes are to each other)–this is defined as (Estrada, 2010).

$$H = \frac{\sum_{i,j\in\Gamma}\left(k_i^{-1/2} - k_j^{-1/2}\right)}{N - 2\sqrt{N-1}},$$

where k_i is the degree of node i, Γ is the edge set.

- Degree Network Centralization (NC) (how much the network is centered around a small number of highly central nodes) as defined by Freeman (1979).

$$NC = \frac{\sum_{i=1}^{N}(k_{\max} - k_i)}{N^2 - 2N + 2},$$

where k_i is the degree of node i, $k_{\max}$ is the maximum node degree in the graph.

- Average Shortest Path Length ($\langle AV\rangle$)
- Exponent when fitting the degree distribution to a power law (α)
- $x_{\min}$ when fitting the degree distribution to a power law ($x_{\min}$)
- Assortativity (A) (Pearson correlation between the rows of the adjacency matrix)

TABLE 1 | Means for the network measures for each dataset. Bold font indicates the mean difference is significant between the AD and controls for that co-occurrence window at $p < 0.05$ level. The parameter *o* refers to the size of the co-occurrence window.

Measure	**o = 2**		**o = 3**		**o = 5**	
	Control	**AD**	**Control**	**AD**	**Control**	**AD**
$< N >$	**64.185**	**53.204**	**64.185**	**53.204**	**64.185**	**53.204**
$< E >$	**151.648**	**124.130**	**206.222**	**168.519**	**281.500**	**226.759**
ED	**0.039**	**0.049**	**0.053**	**0.065**	**0.071**	**0.084**
SL	0.005	0.009	0.016	0.023	0.040	0.039
$< CC >$	0.612	0.601	0.710	0.707	0.792	0.793
D	6.574	6.259	5.037	4.926	4.037	4.167
H	**0.135**	**0.107**	**0.123**	**0.100**	**0.112**	**0.093**
NC	**0.348**	**0.284**	**0.438**	**0.364**	**0.522**	**0.433**
$\langle AV \rangle$	2.724	2.691	2.353	2.305	2.138	2.075
α	5.069	5.477	4.740	4.827	4.171	4.349
$x_{\min}$	4.870	4.815	6.389	5.648	7.630	7.056
A	**–0.159**	**–0.095**	**–0.141**	**–0.075**	**–0.131**	**–0.044**
$k_{nn}\alpha$	**11.614**	**15.575**	13.981	16.409	13.834	18.456

- Exponent when fitting a power law to the average neighbor degree distribution ($k_{nn}\alpha$)

These networks are not connected, and measures that rely on path lengths require modification to be used. In our case the measures that need modifying are the average shortest length path and the diameter. For the average shortest length path, we take the average of all the shortest paths lengths that do exist in the network, discarding those that have infinite length. For the diameter, we take the diameter of the largest component in the graph.

4.2 Results

We show the means of these for each group for a variety of co-occurrence windows (*o*) in **Table 1**, where bold font indicates the mean difference is significant according to a Mann-Whitley test at $p < 0.05$ for that co-occurrence window. Perhaps unsurprisingly, the measures are affected by the size of the co-occurrence window. Increasing the window size increases the number of edges, and by proxy the edge density. As the network becomes more connected, this increases the average clustering coefficient, network centralization and $x_{\min}$ while decreasing the diameter, heterogeneity, average path length and α. We find that there are six measures with significantly different means for all co-occurrence windows, number of nodes (note this is the same for all window sizes), number of edges, edge density, heterogeneity, network centralization and assortativity. There are three other measures that are significant for one window, fraction of self links and $x_{\min}$ for $o = 3$ and $k_{nn}\alpha$ for $o = 2$.

Next we look to explain why these measures might be different. Alzheimer's patients tend to use fewer unique words than controls (Fraser et al., 2016; Orimaye et al., 2018), and tend to repeat words and phrases more frequently than healthy controls. Since unique words correspond to nodes in the graphs, this would explain why controls have a higher number of nodes that those from AD patients, and why the edge density is higher for the AD patients (more edges between a smaller number of nodes). The number of self links should capture word repetitions, and it is higher in the AD networks, but it is notable that it is only significant with a co-occurrence window of 3. For the larger windows there will be more self links overall and proportionally fewer that are due to repetitions, so this could explain why it is not significant for a window of 5.

The AD networks have a lower heterogeneity and a lower network centralization. A lower heterogeneity shows that the degree of the nodes is more equal, while a lower network centralization indicates the network is less orientated around a small number of highly centralized nodes. Furthermore the AD networks are less disassortative than the control networks. A disassortative network is where high degree nodes are connected to low degree nodes, while in an assortative network high degree nodes are connected to other high degree nodes. In general word co-occurrence networks tend to be disassortative (Masucci and Rodgers, 2006; Krishna et al., 2011). We would expect the networks from the AD patients to be smaller, more densely connected and to have a more uniform degree distribution than those from controls, and this seems to be reflected in the graph measures. This would also affect the assortativity of the network–nodes would be less likely to be connected to other nodes of higher degrees, which might indicate greater use of circumlocution in AD networks where disassortativity is reduced.

The average clustering coefficient, diameter and average path length were not significantly different between the AD and controls. We found this surprising as we expected that the average path length and diameter would be shorter for the AD networks as AD patients tend to produce shorter sentences with shorter dependency distances, and the average clustering coefficient larger due to the smaller network size and larger edge density. In fact this is even more surprising as the control networks are larger than the AD networks–so we would expect the diameter and average shortest length path of the control networks to be larger. However, there are disagreements in the literature on whether dependency distance is actually shorter linguistic AD patients' linguistic output (Pakhomov et al., 2011; Fraser et al., 2016; Orimaye et al., 2017), and average path length is not an exact measure of dependency distance. Furthermore, the transcripts of spontaneous speech used in our experiments are quite short, which might have an effect when comparing these results to results from written text, in the context of which the initial claim was made. It should be noted, however, that recent evidence seems to suggest that dependency lengths in spoken language do not differ significantly to those in written language (Kramer, 2021). Other syntax differences discovered were more node level than global (for instance the number of times the participant utters a pronoun and then an auxiliary verb phase) which cannot be picked up by our measures. The other measures which were not significantly different were $x_{\min}$ and α. These describe the degree distribution of the networks. The lack of significant difference in these measures for the majority of the co-occurrence windows indicates that the networks have a similar degree distribution.

Each transcript is also annotated with an MMSE score, and we look at how this is correlated with the network measures using Spearman correlation. A larger MMSE score indicates the participant is less likely to be in the AD group, so we would expect that graph measures which are larger in the controls to be positively correlated with the MMSE score, and those which are

TABLE 2 | Spearman Correlation between network measures and MMSE score. Significant correlations are marked with bold font.

Measure	o = 2	o = 3	o = 5
$\langle N \rangle$	0.187	0.187	0.187
$\langle E \rangle$	0.153	0.154	0.158
ED	**–0.239**	**–0.228**	**–0.203**
SL	**–0.228**	–0.143	0.049
$\langle CC \rangle$	0.185	0.129	0.066
D	0.048	–0.018	–0.155
H	**0.336**	**0.265**	**0.211**
NC	**0.307**	**0.303**	**0.319**
$< AV >$	-0.121	-0.108	-0.085
α	-0.058	-0.075	0.058
$x_{\min}$	0.046	**0.249**	0.117
A	**–0.381**	**–0.421**	**–0.355**
$k_{nn}\alpha$	–0.187	–0.101	–0.117

smaller in the controls to be negatively correlation with the MMSE score. The results are shown in **Table 2**, with significant correlations marked using bold font.

There are five measures with significant correlations with the MMSE score, edge density, heterogeneity, network centralization and assortativity. Edge density and assortativity show a negative relationship with the MMSE score, while heterogeneity and network centralization show a positive relationship. Controls have higher MMSE scores than those with AD, so these results mostly reflect the control/AD differences seen above. There are two measures which have a significant difference in means, but do not have significant correlations, the number of nodes and number of edges. This is quite surprising, as these have been shown to be very good predictors of AD. There is a large amount of variance in the MMSE for both classes, so this could be the reason why the mean difference is significant while the correlation is not.

4.3 Comparison to Shuffled Networks

To understand how successful these networks are in capturing the dynamics of word usage we must compare them to a null model. In this section we create null models by shuffling the order of the words for each transcript and constructing networks from these shuffled transcripts. Previous work comparing shuffled networks to their originals has shown that many of the properties of word networks occur to due the frequency of word use rather than due to word order (Caldeira et al., 2006; Krishna et al., 2011).

We create the shuffled networks by randomizing the order of the words in the document. The end of the sentence marker (usually a full-stop) is treated as a word, so sentence structure is not maintained, but the shuffled documents still have sentences. This is done 50 times for each network and then the mean value for each measure is calculated. These are compared to the originals. This allows us to see which structures of the network are due to the frequency of word occurrence and which are due to the specific word order. We show the results of this in **Table 3**. Again we use a Mann-Whitley test at $p < 0.05$ to test for means that are significantly different.

Some of the measures are obviously more influenced by the number of words than their order - for instance the number of nodes, number of edges and edge density, and we can see these are not significantly different between the real and shuffled networks for any value. Only the average clustering coefficient, the number of self links and $x_{\min}$ are significantly different between the real and shuffled for all the networks. Assortativity is also different for all the control networks, but only for the co-occurrence window of two for the AD networks.

TABLE 3 | Comparison of the network measures for the shuffled networks and the real ones. Significant differences are marked with bold font.

o = 2

Measure	Control		AD	
	Real	Shuffled	Real	Shuffled
$< N >$	64.185	64.006	53.204	53.259
$< E >$	151.648	156.699	124.130	136.437
ED	0.039	0.041	0.049	0.052
SL	**0.005**	**0.035**	**0.009**	**0.041**
$< CC >$	**0.612**	**0.583**	**0.601**	**0.571**
D	6.574	6.316	6.259	6.052
H	0.135	0.134	0.107	0.121
NC	0.348	0.359	0.284	0.306
$< AV >$	2.724	2.855	2.691	2.699
α	5.069	4.983	5.477	5.629
$x_{\min}$	**4.870**	**5.429**	**4.815**	**5.253**
A	**–0.159**	**–0.108**	**–0.095**	**–0.088**
$k_{nn}\alpha$	**11.614**	**11.090**	15.575	13.362

o = 3

Measure	Control		AD	
	Real	Shuffled	Real	Shuffled
$\langle N \rangle$	64.185	63.983	53.204	53.220
$< E >$	206.222	212.448	168.519	183.404
ED	0.053	0.055	0.065	0.069
SL	**0.016**	**0.045**	**0.023**	**0.051**
$\langle CC \rangle$	**0.710**	**0.679**	**0.707**	**0.669**
D	5.037	5.029	4.926	4.852
H	0.123	0.124	**0.100**	0.113
NC	0.438	0.436	0.364	0.373
$\langle AV \rangle$	2.353	2.505	2.305	2.349
α	4.740	4.616	**4.827**	**5.276**
$x_{\min}$	**6.389**	**7.111**	**5.648**	**6.822**
A	**–0.141**	**–0.110**	–0.075	–0.082
$k_{nn}\alpha$	13.981	12.655	16.409	16.283

o = 5

Measure	Control		AD	
	Real	Shuffled	Real	Shuffled
$< N >$	64.185	64.005	53.204	53.206
$< E >$	281.500	297.417	226.759	252.929
ED	0.071	0.076	**0.084**	**0.093**
SL	**0.040**	**0.056**	**0.039**	**0.065**
$\langle CC \rangle$	**0.792**	**0.759**	**0.793**	**0.752**
D	4.037	4.083	4.167	4.067
H	0.112	0.112	0.093	0.104
NC	0.522	0.522	0.433	0.445
$< AV >$	2.138	2.257	2.075	2.132
α	**4.171**	**4.403**	**4.349**	**5.113**
$x_{\min}$	**7.630**	**8.929**	**7.056**	**8.975**
A	**–0.131**	**–0.105**	–0.044	–0.064
$k_{nn}\alpha$	13.834	14.539	**18.456**	**19.742**

For the average clustering coefficient, this significant difference is explained by the fact that words co-occur more than would be expected due to random chance. Shuffling destroys this structure, and therefore reduces the clustering coefficient in all of the networks. The difference in self links should also be caused by a similar situation–this measure is clearly influenced by word order, and so should change when this is destroyed. Shuffling also changes the degree structure of the networks, causing changes in the value calculated for $x_{\min}$.

In the previous section we found that there are six network measures that differ between the AD and controls for all co-occurrence windows: number of edges, number of nodes, edge density, heterogeneity, network centralization and assortativity. However out of all of these only assortativity differs between the shuffled and original networks. From a purely network based perspective, it would seem reasonable that assortativity would change between the shuffled and original networks–again we are destroying the co-occurrence structure. Previous work (Krishna et al., 2011) has also confirmed this. However what is surprising is that the difference is significant for controls for all co-occurrence windows, but only for $o = 2$ for AD patients. This indicates the AD networks look more random than those from controls.

Previous work has shown that AD patients tend to use more generic terms on picture description tasks than healthy controls, and that the healthy controls use more low frequency content bearing words (Garrard et al., 2014). These two factors help to explain why both heterogeneity and network centralization differ between the AD and controls, but not between the shuffled networks–AD patients will tend to use a smaller set of words, but use each of these words more frequently compared to healthy controls. This indicates that word frequency has the largest impact on the structure of the networks, and we would therefore conclude that word frequency statistics alone would still provide a good feature set to distinguish between the two classes of networks.

5 TRANSCRIPT CLASSIFICATION

5.1 Method

We are interested in methods for automatic classification of networks into control or AD. This can be done in a variety of ways, with previous work on work co-occurrence networks often using the network measures above as input into a classifier. However there has been a great deal of work in the area of graph classification in the past few years, with many methods being proposed. Generally these methods fall into one of two broad categories: embedding or kernel methods. An embedding method reduces a graph to a vector, while a kernel method learns some kind of similarity measure between graphs and calculates the Gram matrix from this (Kriege et al., 2020). In addition to using the network measures mentioned in the previous section, we also use the spectral features (SF) embedding method created by de Lara and Pineau (2018). This method is based on analyzing the spectrum of the graph's Laplacian in order to extract a feature vector for the classification algorithm. Firstly we calculate the normalized graph Laplacian

$$L = I - D^{-1/2} A D^{-1/2}, \tag{1}$$

where D is the degree matrix, A is the adjacency matrix of the graph and I is the identity matrix. The input into the classifier is then the k smallest eigenvalues of the Laplacian in ascending order

$$X = (\lambda_1, \lambda_2, \ldots \lambda_k). \tag{2}$$

The authors claim that this is similar to classifying a melody by its lowest fundamental frequencies. A deeper explanation of the method is undertaken by Pineau (2019). A larger vector will capture more of the dynamics of the graph, but will also be more prone to overfitting. Since we are not aware of an objective method of selecting k, we experiment with the size of the vector, running for 5, 10, 15, 20 and 50.

A particular emphasis here is that we are not using the word labels in this classification task, but purely the structure of the networks. As mentioned in the previous section, many of the network measures that differ between the AD and controls do not vary between the shuffled and original networks, indicating that many of the differences are due to word frequency usage alone. With this in mind, we use a unigram based method to provide a baseline comparison as to how much word frequency alone can be used to differentiate between the two classes. Here we take the number of different words used and total number of words in the transcript, and then the mean, standard deviation, skew and kurtosis of the distribution of unigrams in the transcript. When creating the distributions of unigrams we only consider the unigrams in the specific transcript. This ensures that we do not leak information across transcripts, and to provide a fairer comparison to the graph measures, as one of the advantages of these graph approaches is that we do not need to consider which words occur in other transcripts.

We use logistic regression (LR), a linear kernel (LSVM), a radial basis function (RBF) kernel (RSVM), and a random forest (RF) to classify the networks. The input variables are standardized to have a mean of 0 and a standard deviation of 1. C for the SVMs is set to 1, and the logistic regression is L_2 regularized, with a regularization parameter of 0.5. γ for the RBF kernel is set to $1/p$.

5.2 Results

Firstly we evaluate our approach using leave one out cross-validation (LOOCV) on the training set, and the results are shown in **Figure 2**. From this we can see that it is possible to distinguish between the co-occurrence networks. We have the highest success at the smallest co-occurrence window of $o = 2$ using a linear SVM, with a classification accuracy of 71.3% using the graph measures. Using the graph measures has a higher overall success rate than using the embedding method for every size of co-occurrence window. However the unigram method actually outperforms the network based methods, achieving a maximum accuracy of 73.1% using a linear SVM.

There is not one particular classification algorithm that consistently outperforms the other, with the logistic regression, random forest and linear SVM all having the highest classification accuracy for different co-occurrence window sizes. The co-occurrence window size obviously does not affect the unigram

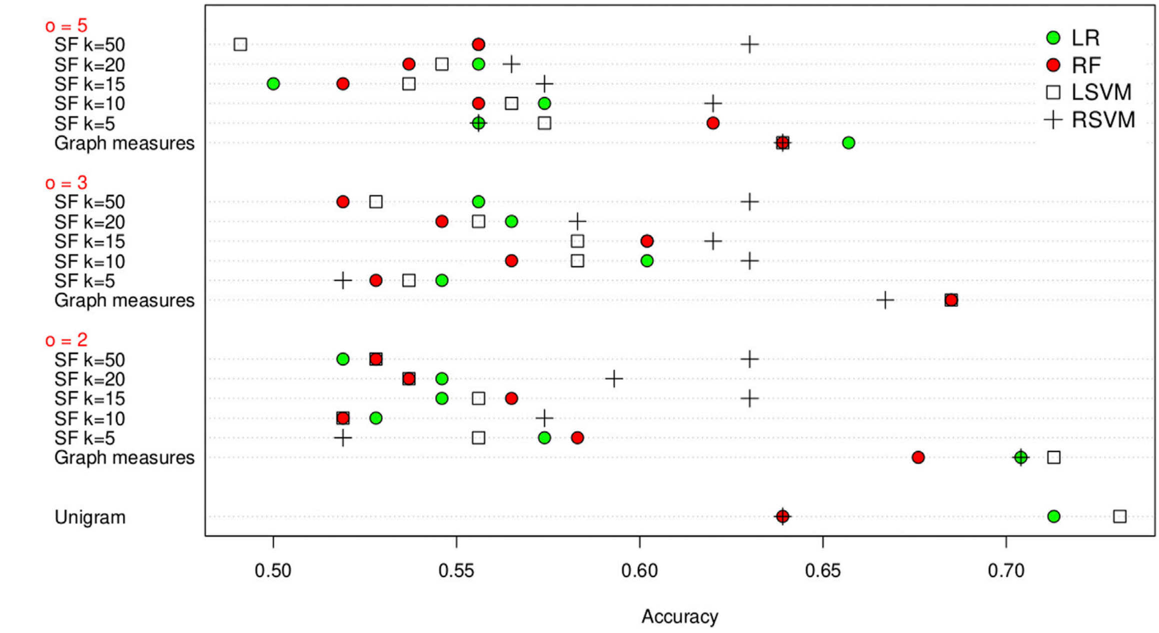


FIGURE 2 | Classification accuracy of the unigram, graph measures and SF feature extraction methods on the training set using leave one out cross validation, grouped by co-occurrence window size ($o \in \{2, 3, 5\}$). The best result of 71.3% is achieved using a linear SVM on the graph measures with a co-occurrence window of 2 ($o = 2$).

methods. To further evaluate this we use a Wilcoxon signed-rank test to look if the differences in classification accuracy are significant. Again we take $p < 0.05$ as a significant difference. To start with we compare the unigram and graph measure feature sets. The only significant difference between them is for the LSVM classifier at $o = 5$ (which is the best performing unigram combination against the worst performing graph measures combination), indicating their performance is broadly similar.

Next we compare how the choice of k affects the results for SF. There are some significant differences with the RSVM for $k = 5$ with a co-occurrence window of three performing significantly worse than the same classifier for the rest of the values of k, and the different between the RSVM for $k = 50, o = 5$ performing significantly better than $k = 20, o = 5$. The rest are not significant, indicating that in general, the choice of k is not particularly important. Comparing the results between the different co-occurrence windows, we find no significant differences for the graph measures. This implies that the choice of co-occurrence window is not particularly important. This again confirms that word frequency seems more important than word co-occurrence.

Looking at the same comparison for SF, there are three classifier/feature sets with a significant difference, logistic regression with $k = 10$ between $o = 2$ and $o = 3$, logistic regression with $k = 15$ between $o = 3$ and $o = 5$ and RSVM with $k = 15$ between $o = 3$ and $o = 5$. Again with the small number of significant differences, we would conclude than the co-occurrence window choice does not particularly affect the SF method.

TABLE 4 | Classification accuracy of the best performing embedding/classifier combination from the training set on the test set. We choose the three best performing graph measure approaches, the two best SF approaches and the best unigram approach for comparison. There is a decrease in performance in general compared to the training set, but the best performing approach (graph measures with a RSVM classifier) achieves a classification accuracy of 66.7%.

Method	o	Accuracy
GM + RF	3	0.583
GM + LSVM	2	0.625
GM + RSVM	2	0.667
SF $k = 5$ + RF	5	0.583
SF $k = 10$ + RSVM	5	0.542
Unigram + RF		0.646

As previously mentioned, the ADReSS dataset contains a pre-tagged test set. Next we look at our success in distinguishing between the AD and control transcripts in the test set. We choose the three best performing classifier/co-occurrence window combinations on the training set for the graph measures, and the two best performing SF methods, in the manner of the ADReSS challenge. The results are shown in **Table 4** and confusion matrices in **Figure 3**. For the combinations that perform the best on the training set, the maximum accuracy achieved is 66.7% using a RSVM classifier with graph measures with graphs that have a co-occurrence window of 2. Bar this outlier though, the accuracy in general has dropped when compared to the results of the leave one out cross-validation on the training set. However, as the test set consists of a very small

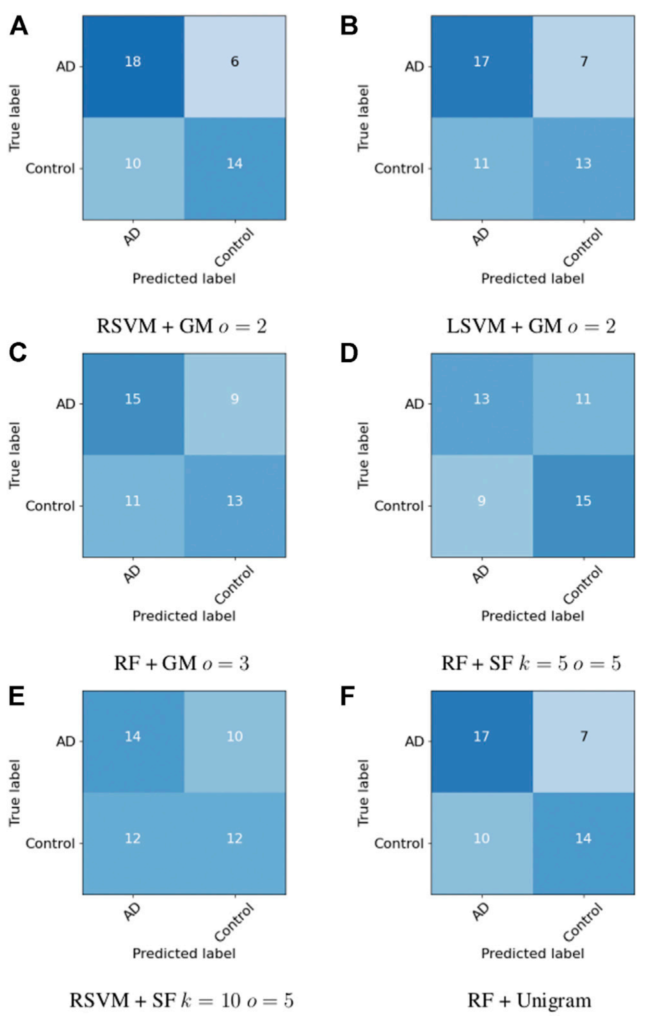


FIGURE 3 | Confusion matrices for the test set.

sample, it is likely that the reported LOOCV accuracy gives a more realistic assessment of the methods we compared.

Ignoring performance on the training set and purely taking the classifier/feature combination that has the highest performance, we can achieve an accuracy of 75% using a random forest with a co-occurrence window of 2. However since this method did not perform so well on the training set, it is difficult to claim that this is an accurate and reportable classification accuracy.

6 MMSE PREDICTION

In this section we focus on using the co-occurrence networks to predict the MMSE score for a participant. As with the classification in **Section 5**, we use the network measures, the SF graph embedding method and the unigram method as features. We choose a set of regression methods analogous to the classification methods chosen above, in this case Linear Regression, Random Forest Regression, and Support Vector Regression with two kernels, a linear kernel and RBF kernel. The input into the regression methods is again standardized so each feature has a mean of 0 and a standard deviation of 1. The predictions are evaluated using root mean squared error (RMSE). C for the SVMs is set to 1, and the linear regression is L_2 regularized, with a regularization parameter of 0.5. γ for the RBF kernel is set to $1/p$.

As before, we firstly evaluate the method using LOOCV on the training set. The results are shown in **Figure 4**. Using a linear regression method with the graph measures appears to obtain the best result (i.e. lowest RMSE), with a RMSE of 4.799 for a co-occurrence window of 2. Again the graph measures seem to give the best results. Following LOOCV, we predict the MMSE of the test set transcripts. As before we take the five embedding/regression combinations that perform the best on the training set and evaluate their performance on the test set. The results are shown in **Table 5**. Again we do see a decrease in the success of the methods compared to the leave one out evaluation on the training set, with an increase in the RMSE. This time the unigram measures actually give the lowest RMSE, at 5.468, by using linear regression. The best performing graph method uses the graph measures and a random forest regressor with a co-occurrence window of 3, achieving an RMSE of 5.675.

As RMSE values can be difficult to interpret in isolation, we also use the predicted MMSE value to assign the participant as AD or control (a value above 23 indicates a control). The results of this are shown in **Table 6**. This approach achieves a maximum accuracy of 75% for the graph measures, 64.6% for the unigram methods, and 58.3% for the SF methods. To give a reference, if we use the actual MMSE values for AD prediction, we get an accuracy of 87.5%.

7 DISCUSSION AND CONCLUSION

In this paper we have constructed word co-occurrence networks using transcript data from both controls and Alzheimer's patients on a picture description task. With these networks we have analyzed some measures of their structure, and used some embedding methods to enable classification of the networks and to predict the MMSE score from the transcript.

Using a Mann-Whitney test we find that there are six measures that have significantly different means between the networks, number of nodes, number of edges, edge density, heterogeneity, network centralization and assortativity. Some of this difference can be explained by previous work in the literature, for instance that AD patients tend to produce fewer unique words and repeat themselves more. Most of these measures also show significant correlation with the MMSE score of the participant. However, many of the measures that differ between the AD and control networks do not differ between the shuffled and original networks. This is unfortunately one of the challenges of using global measures on co-occurrence networks, that in fact many of their properties come from word frequency rather than co-occurrence.

We then looked at classifying the graphs into control or AD using the set of graph measures, and the graph embedding method, SF. Since many of the graph properties come from word frequency and not co-occurrence, we create a baseline feature set using the

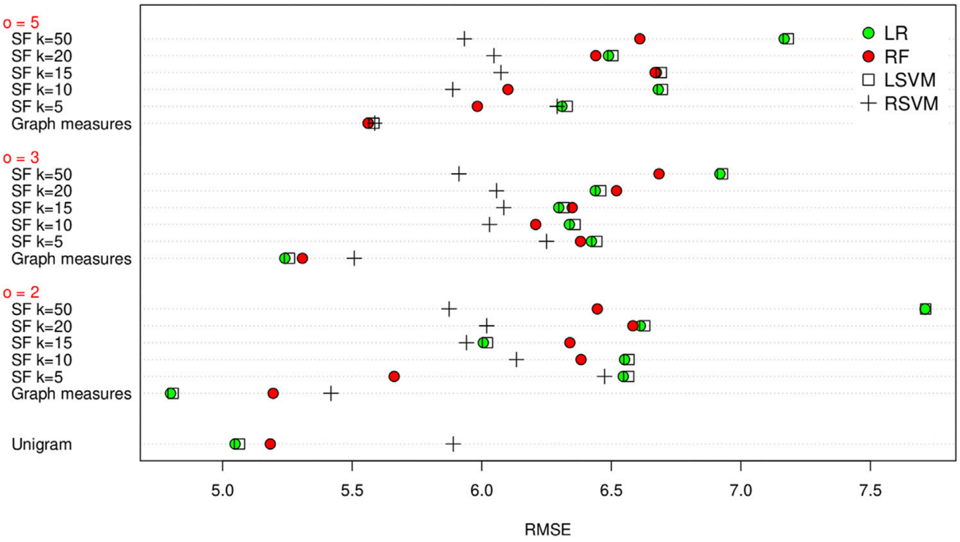


FIGURE 4 | RMSE of the unigram, graph measures and SF methods for leave one out cross validation on the training set, grouped by co-occurrence window size ($o \in \{2, 3, 5\}$).

first four moments of the unigram distribution, plus the total number of unigrams and the number of unique unigrams. We evaluate our success in this firstly by using leave one out cross validation on the training set, and then by using the held back test set from the ADReSS challenge.

In general we find it is possible to classify the networks into control or AD, and that the highest accuracy on the training set is achieved using graph measures and a Linear SVM at 71.3%. For the test set, the highest accuracy achieved is 66.7%, using a RSVM classifier with a co-occurrence window of 2. Using the graph measures gives a higher accuracy than using the SF method, but out of the four classifiers we use, three (Logistic Regression, Random Forest and Linear SVM) have the highest performance at one particular point, making it difficult to recommend the use of one in particular. The same applies to the choice of co-occurrence window. We also find that using the unigram gives fairly comparable results to the graph measures, further indicating that global measures on these word co-occurrence networks mostly reflect word frequency rather than word co-occurrence.

In a similar manner to the graph classification, we also look at predicting the MMSE score from the transcripts. We use the same evaluation methods, leave one out cross validation on the training set, and using the held back test set. On the training set we achieve a minimum RMSE of 4.799 using linear regression and the graph measures with a co-occurrence window of 2, and on the test set we achieve a minimum RMSE of 5.675 using linear regression and the graph measures with a co-occurrence window of 3. Here the unigram methods perform notably better, achieving an RMSE of 5.468 using linear regression. Again the SF method performed poorly compared to the other methods, achieving a maximum accuracy of 6.535 on the test set.

In our work, we have found that using simple unigram measures outperforms using more complex graph based measures which should take co-occurrence into account. However, looking at the features that previous work has found to be useful in distinguishing between AD and control patients, it could be that the measures we have chosen cannot capture these differences with a great deal with success. Combined with previous work showing that global network measures on word co-occurrence networks struggle to capture word order, we would suggest that future work either relies node level measures, or devises novel global measures that can capture word order. We also note that our network-based approach performed comparably to the

TABLE 5 | MMSE of the best performing embedding/regression combination from the training set on the test set. We choose the three best performing graph measure approaches, the two best SF approaches and the best unigram method.

Method	o	RMSE
GM + LR	2	6.154
GM + LSVM	2	6.159
GM + RF	2	5.675
SF k = 10 + RF	3	6.535
SF k = 10 + RSVM	5	6.535
Unigram + LR		5.468

TABLE 6 | Accuracy of the best performing regression classifier/embedding methods if we use the predicted MMSE score to predict a transcript as AD or control.

Method	o	RMSE
GM + LR	2	0.667
GM + LSVM	2	0.667
GM + RF	2	0.750
SF k = 10 + RF	3	0.542
SF k = 10 + RSVM	5	0.521
Unigram + LR		0.583

ADReSS baseline, with scores of 5.68 vs 5.20 RMSE for regression, and 66.7% vs 75.0% for classification (Luz et al., 2020). However, none of the participants employed a network based approach.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://dementia.talkbank.org/. Instructions on how to acquire the dataset are available at https://edin.ac/375QRNI. The source code for the experiments described in this paper is available at https://git.ecdf.ed.ac.uk/tmilling/analysis-and-classification-of-word-co-occurrence-networks.

AUTHOR CONTRIBUTIONS

TM and SL conceived and designed the experiments and analysis. SL prepared the dataset. TM implemented the network-generation and feature extraction algorithms, performed analysis and drafted the manuscript. Both authors contributed to the final version of the manuscript.

FUNDING

This work funded by the European Union's Horizon 2020 research and~innovation program, under the grant agreement No 769661, toward the SAAM project.

REFERENCES

Akimushkin, C., Amancio, D. R., and Oliveira, O. N. (2017). Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks. *PloS one* 12, e0170527. doi:10.1371/journal.pone.0170527

Alberdi, A., Weakley, A., Schmitter-Edgecombe, M., Cook, D. J., Aztiria, A., Basarab, A., et al. (2018). Smart Home-Based Prediction of Multidomain Symptoms Related to Alzheimer's Disease. *IEEE J. Biomed. Health Inform.* 22, 1720–1731. doi:10.1109/jbhi.2018.2798062

Amancio, D. R., Antiqueira, L., Pardo, T. A. S., da F. COSTACosta, L. L., Oliveira, O. N., and Nunes, M. G. V. (2008). Complex Networks Analysis of Manual and Machine Translations. *Int. J. Mod. Phys. C* 19, 583–598. doi:10.1142/S0129183108012285

Amancio, D. R. (2015). Comparing the Topological Properties of Real and Artificially Generated Scientific Manuscripts. *Scientometrics* 105, 1763–1779. doi:10.1007/s11192-015-1637-z

Antiqueira, L., Pardo, T. A. S., Nunes, M. d. G. V., and Oliveira, O. N. (2007). Some Issues on Complex Networks for Author Characterization. Inteligencia Artificial. *Revista Iberoamericana de Inteligencia Artif.* 11, 51–58. doi:10.4114/ia.v11i36.891

Barrenechea, J., Bullmore, E., and Plenz, D. (2014). Powerlaw: A python Package for Analysis of Heavy-Tailed Distributions. *PLoS One* 9, e85777–e857711. doi:10.1371/journal.pone.0085777

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The Natural History of Alzheimer's Disease. *Arch. Neurol.* 51, 585–594. doi:10.1001/archneur.1994.00540180063015

Bougouin, A., Boudin, F., and Daille, B. (2013). "TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction," in Proceedings of the Sixth International Joint Conference on Natural Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, 543–551.

Caldeira, S. M. G., Petit Lobão, T. C., Andrade, R. F. S., Neme, A., and Miranda, J. G. V. (2006). The Network of Concepts in Written Texts. *Eur. Phys. J. B* 49, 523–529. doi:10.1140/epjb/e2006-00091-3

Cancho, R. F. i., and Solé, R. V. (2001). The Small World of Human Language. *Proc. R. Soc. Lond. B* 268, 2261–2265. doi:10.1098/rspb.2001.1800

Cong, J., and Liu, H. (2014). Approaching Human Language with Complex Networks. *Phys. Life Rev.* 11, 598–618. doi:10.1016/j.plrev.2014.04.004

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *J. Alzheimer's Dis.* 78, 1547–1574. doi:10.3233/JAD-200888

de Lara, N., and Pineau, E. (2018). A Simple Baseline Algorithm for Graph Classification. Available at: https://arxiv.org/abs/1810.09155.

Estrada, E. (2010). Quantifying Network Heterogeneity. *Phys. Rev. E* 82, 066102. doi:10.1103/physreve.82.066102

Florescu, C., and Caragea, C. (2017). "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vancouver, Canada: Association for Computational Linguistics) 1, 1105–1115. doi:10.18653/v1/P17-1102

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *J. Alzheimers Dis.* 49, 407–2210. doi:10.3233/JAD-150520

Freeman, L. C. (1979). Centrality in Social Networks I: Conceptual Clarification. *Social Networks* 1, 215–239. doi:10.1016/0378-8733(78)90021-7

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2005). The Effects of Very Early Alzheimer's Disease on the Characteristics of Writing by a Renowned Author. *Brain* 128, 250–6010. doi:10.1093/brain/awh341

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., and Gorno-Tempini, M. L. (2014). Machine Learning Approaches to Diagnosis and Laterality Effects in Semantic Dementia Discourse. *Cortex* 55, 122–129. doi:10.1016/j.cortex.2013.05.008

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). "Exploring Network Structure, Dynamics, and Function Using Networkx," in Proceedings of the 7th Python in Science Conference. Pasadena, CA, 11–15. doi:10.25080/issn.2575-9752

Haider, F., De La Fuente, S., and Luz, S. (2020). An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE J. Sel. Top. Signal. Process.* 14, 272, 281. doi:10.1109/JSTSP.2019.2955022

Hassan, S., Mihalcea, R., and Banea, C. (2007). Random Walk Term Weighting for Improved Text Classification. *Int. J. Semantic Comput.* 01, 421–439. doi:10.1142/s1793351x07000263

Hunter, J. D. (2007). Matplotlib: A 2d Graphics Environment. *Comput. Sci. Eng.* 9, 90–95. doi:10.1109/MCSE.2007.55

Kramer, A. (2021). Dependency Lengths in Speech and Writing: A Cross-Linguistic Comparison via Youdepp, a Pipeline for Scraping and Parsing Youtube Captions. *Proc. Soc. Comput. Linguistics* 4, 359–365.

Kriege, N. M., Johansson, F. D., and Morris, C. (2020). A Survey on Graph Kernels. *Appl. Netw. Sci.* 5, 1–42. doi:10.1007/s41109-019-0195-3

Krishna, M., Hassan, A., Liu, Y., and Radev, D. (2011). The Effect of Linguistic Constraints on the Large Scale Organization of Language. Available at: https://arxiv.org/abs/1102.2831.

Lee, J. L., Burkholder, R., Flinn, G. B., and Coppess, E. R. (2016). Working with CHAT Transcripts in Python. *Tech. Rep. TR-2016-02*. Department of Computer Science, University of Chicago.

Liu, H., and Cong, J. (2013). Language Clustering with Word Co-occurrence Networks Based on Parallel Texts. *Chin. Sci. Bull.* 58, 1139–1144. doi:10.1007/s11434-013-5711-8

Loper, E., and Bird, S. (2002). "NLTK: The Natural Language Toolkit," in Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - 1. Philadelphia, PA: Association for Computational Linguistics, 63–70. doi:10.3115/1118108.1118117

Luz, S., de la Fuente, S., and Albert, P. (2018). A Method for Analysis of Patient Speech in Dialogue for Dementia Detection. Available at: http://arxiv.org/abs/1811.09919.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in Proceedings of INTERSPEECH 2020. doi:10.21437/interspeech.2020-2571

MacWhinney, B. (2019). Understanding Spoken Language through Talkbank. *Behav. Res.* 51, 1919–1927. doi:10.3758/s13428-018-1174-9

Masucci, A. P., and Rodgers, G. J. (2006). Network Properties of Written Human Language. *Phys. Rev. E* 74, 026102. doi:10.1103/PhysRevE.74.026102

McKinney, W. (2010). "Data Structures for Statistical Computing in python," in Proceedings of the 9th Python in Science Conference. Editors S. van der Walt and J. Millman, 51–56. doi:10.25080/issn.2575-9752

Mehri, A., Darooneh, A. H., and Shariati, A. (2012). The Complex Networks Approach for Authorship Attribution of Books. *Physica A: Stat. Mech. its Appl.* 391, 2429–2437. doi:10.1016/j.physa.2011.12.011

Mihalcea, R., and Tarau, P. (2004). "Textrank: Bringing Order into Text," in Proceedings of the 2004 conference on empirical methods in natural language processing. 404–411. doi:10.3115/1220355.1220517

Oliphant, T. E. (2006). *A guide to NumPy*. Scotts Valley, CA: CreateSpace..

Orimaye, S. O., Wong, J. S.-M., and Wong, C. P. (2018). Deep Language Space Neural Network for Classifying Mild Cognitive Impairment and Alzheimer-type Dementia. *PLoS One* 13, e0205636–15. doi:10.1371/journal.pone.0205636

Orimaye, S. O., Wong, J. S., Golden, K. J., Wong, C. P., and Soyiri, I. N. (2017). Predicting Probable Alzheimer's Disease Using Linguistic Deficits and Biomarkers. *BMC bioinformatics* 18, 34. doi:10.1186/s12859-016-1456-0

Pakhomov, S., Chacon, D., Wicklund, M., and Gundel, J. (2011). Computerized Assessment of Syntactic Complexity in Alzheimer's Disease: a Case Study of Iris Murdoch's Writing. *Behav. Res.* 43, 136–144. doi:10.3758/s13428-010-0037-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in python. *J. Machine Learn. Res.* 12, 2825–2830.

Pineau, E. (2019). Using Laplacian Spectrum as Graph Feature Representation. Available at: http://arvix.org/abs/1912.00735.

Rousseau, F., Kiagias, E., and Vazirgiannis, M. (2015). "Text Categorization as a Graph Classification Problem," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Beijing, China: Association for Computational Linguistics)) 1, 1702–1712. doi:10.3115/v1/P15-1164

Rozemberczki, B., Kiss, O., and Sarkar, R. (2020). An Api Oriented Open-Source python Framework for Unsupervised Learning on Graphs. Available at: http://arvix.org/abs/2003.04819..

Santos, L. B. d., Corrêa, E. A., Oliveira, O. N., Amancio, D. R., Mansur, L. L., and Aluísio, S. M. (2017). Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. Available at: https://arxiv.org/abs/1704.08088..

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Slegers, A., Filiou, R.-P., Montembeault, M., and Brambati, S. M. (2018). Connected Speech Features from Picture Description in Alzheimer's Disease: A Systematic Review. *Jad* 65, 519–542. doi:10.3233/jad-170881

Wankerl, S., Nöth, E., and Evert, S. (2017). An N-Gram Based Approach to the Automatic Diagnosis of Alzheimer's Disease from Spoken Language. *INTERSPEECH*, 3162–3166. doi:10.21437/Interspeech.2017-1572

Yan, X., and Han, J. (2002). "Gspan: Graph-Based Substructure Pattern Mining," in IEEE International Conference on Data MiningProceedings. IEEE, 721–724.

Conflict of Interest: The 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.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 12 May 2021
doi: 10.3389/fcomp.2021.624683

Exploring Deep Transfer Learning Techniques for Alzheimer's Dementia Detection

Youxiang Zhu[1], Xiaohui Liang[1], John A. Batsis[2] and Robert M. Roth[3]*

[1] Computer Science, University of Massachusetts Boston, Boston, MA, United States, [2] School of Medicine, University of North Carolina, Chapel Hill, NC, United States, [3] Geisel School of Medicine at Dartmouth, Lebanon, NH, United States

Examination of speech datasets for detecting dementia, collected via various speech tasks, has revealed links between speech and cognitive abilities. However, the speech dataset available for this research is extremely limited because the collection process of speech and baseline data from patients with dementia in clinical settings is expensive. In this paper, we study the spontaneous speech dataset from a recent ADReSS challenge, a Cookie Theft Picture (CTP) dataset with balanced groups of participants in age, gender, and cognitive status. We explore state-of-the-art deep transfer learning techniques from image, audio, speech, and language domains. We envision that one advantage of transfer learning is to eliminate the design of handcrafted features based on the tasks and datasets. Transfer learning further mitigates the limited dementia-relevant speech data problem by inheriting knowledge from similar but much larger datasets. Specifically, we built a variety of transfer learning models using commonly employed MobileNet (image), YAMNet (audio), Mockingjay (speech), and BERT (text) models. Results indicated that the transfer learning models of text data showed significantly better performance than those of audio data. Performance gains of the text models may be due to the high similarity between the pre-training text dataset and the CTP text dataset. Our multi-modal transfer learning introduced a slight improvement in accuracy, demonstrating that audio and text data provide limited complementary information. Multi-task transfer learning resulted in limited improvements in classification and a negative impact in regression. By analyzing the meaning behind the Alzheimer's disease (AD)/non-AD labels and Mini-Mental State Examination (MMSE) scores, we observed that the inconsistency between labels and scores could limit the performance of the multi-task learning, especially when the outputs of the single-task models are highly consistent with the corresponding labels/scores. In sum, we conducted a large comparative analysis of varying transfer learning models focusing less on model customization but more on pre-trained models and pre-training datasets. We revealed insightful relations among models, data types, and data labels in this research area.

Keywords: Alzheimer's disease, early detection, spontaneous speech, deep learning, transfer learning

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Rubén San-Segundo,
Polytechnic University of Madrid,
Spain
Sena Büsra Yengeç Tasdemir,
Abdullah Gül University, Turkey
Kevin El Haddad,
University of Mons, Belgium

****Correspondence:***
Xiaohui Liang
Xiaohui.Liang@umb.edu

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *31 October 2020*
Accepted: *11 March 2021*
Published: *12 May 2021*

Citation:
Zhu Y, Liang X, Batsis JA and
Roth RM (2021) Exploring Deep
Transfer Learning Techniques for
Alzheimer's Dementia Detection.
Front. Comput. Sci. 3:624683.
doi: 10.3389/fcomp.2021.624683

1. INTRODUCTION

The number of patients with Alzheimer's Disease (AD) over the age of 65 is expected to reach 13.8 million by 2050, leading to a huge demand on the public health system (Alzheimer's Association, 2020). While there is no proven effective treatment on AD, considerable effort has been put forth into early detection of AD, such that interventions can be implemented at that stage. Screening measures, neuropsychological assessments, and neuroimaging scans are not pragmatic, cost-, or time-efficient approaches for widespread use.

Expressive language impairment is common in AD, such as reduced verbal fluency and syntactic complexity, increased semantic and lexical errors, generating more high-frequency words and shorter utterances, and abnormalities in semantic content (Sajjadi et al., 2012; Fraser et al., 2016; Boschi et al., 2017; Mueller et al., 2018a). Expressive language impairment has also been observed in patients with Mild Cognitive Impairment (MCI), a population at high risk for the development of AD (Mueller et al., 2018b; Kim et al., 2019; Themistocleous et al., 2020). Furthermore, recent meta-analytic and systematic reviews have found that measures of expressive language contribute to the prediction of progression from MCI to AD (Belleville et al., 2017; Prado et al., 2019).

Researchers have explored spontaneous speech as a means of practical and low-cost early detection of dementia symptoms. Pitt Corpus (Becker et al., 1994), one of the large speech datasets, includes spontaneous speech obtained from a Cookie Theft Picture (CTP) description task. Since then, the CTP task has become popular in dementia research and it has been further explored with computerized agents to automate and mobilize the speech collection process (Mirheidari et al., 2017, 2019b) and in other languages including Mandarin (Chien et al., 2019; Wang et al., 2019a), German (Sattler et al., 2015), and Swedish (Fraser et al., 2019b). Other spontaneous speech datasets for dementia research include those collected from film-recall tasks (Tóth et al., 2018), story-retelling tasks (Fraser et al., 2013), map-based tasks (de la Fuente Garcia et al., 2019), and human conversations (Mirheidari et al., 2019a). While a number of studies have investigated speech and language features and machine learning techniques for the detection of AD and MCI, this research field still lacks balanced and standardized datasets on which these different approaches can be systematically and fairly evaluated.

Speech datasets available for dementia research are often small. As shown in **Table 1**, if we consider AD and non-AD as two classes, the numbers of user-samples in each class are in the hundreds. In the past few years, researchers have explored handcrafted features and machine learning algorithms with these datasets for building classification and regression models. Mueller et al. (2018a) published a survey to show effective linguistic features including semantic content, syntax and morphology, pragmatic language, discourse fluency, speech rate, and speech monitoring. The linguistic features were often identified manually, and the analysis methods were complex and highly task and data dependent. Croisile et al. (1996) manually extracted 23 information units from the picture using language knowledge that were effective in dementia detection. Fraser et al. (2019a) developed an auto-generation process of information units for the analysis. Yancheva and Rudzicz (2016) and Fraser et al. (2019b) further proposed to auto-generate topic models that can recall 97% of the human-annotated information units. Similarly, the acoustic-based analysis was started with pre-defined features and recently automated with computational models. Hoffmann et al. (2010) considered acoustic features for each utterance. Fraser et al. (2013) evaluated the statistical significance of pause and word acoustic features. Tóth et al. (2015) considered four descriptors for silent/filled pauses and phonemes. Gosztolya et al. (2016) and Tóth et al. (2018) implemented a customized automatic speech recognition (ASR) and automatic feature selection for phones, boundaries, and filled pauses. Haider et al. (2019), Luz et al. (2020) proposed an automatic acoustic analysis approach using the paralinguistic acoustic features of audio segments. However, the performance results of handcrafted features and customized machine learning algorithms are highly dependent on the tasks and datasets. In 2020, the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge became the first shared-task event focused on AD detection (Luz et al., 2020). The ADReSS organizers pre-processed the CTP dataset of the Pitt Corpus and provided the same dataset to the challenge participants, enabling a fair competition. The techniques and results in this paper will strictly follow the guideline of the ADReSS Challenge.

In recent years, *transfer learning techniques* have significantly advanced the research on Image Recognition (IR), Automatic Speech Recognition (ASR), and Natural Language Processing (NLP). Transfer learning focuses on storing knowledge gained from an easy-to-obtain large-sized dataset from a general task and applying the knowledge to a downstream task where the downstream data is limited. A typical transfer learning model incorporates a pre-trained model as its backbone and is later customized for the downstream task. The pre-training process is computationally intensive and requires a dataset of sufficient size. Different pre-trained models result in different performances as they inherit different knowledge from the pre-training datasets. It is commonly believed that the higher similarity between the pre-training and downstream datasets results in better performance of the downstream task. In addition to the selection of an effective pre-trained model, the customization of the transfer

Abbreviations: ADRD, Alzheimer's Disease and Related Dementias; AD, Alzheimer's Disease; MCI, Mild Cognitive Impairment; HC, Health Control; WLS,Wisconsin Longitudinal Study; CTP, Cookie Theft Picture; IVA, Intelligent Virtual Agent; IU, Information Units; MFCC, Mel Frequency Cepstral Coefficient; LLDs, Low-Level Descriptors; LSP, Line Spectral Pair; AOI, Area of Interest; ASR, Automatic Speech Recognition; ML,Machine Learning; MMSE,Mini-Mental State Examination; MoCA, Montreal Cognitive Assessment; GDS, Geriatric Depression Scale; GAI, Geriatric Anxiety Inventory; SVM, Support Vector Machine; PCA, Principal Component Analysis; DNN, Deep Neural Network; MECSD, Mandarin Elderly Cognitive Speech Database; LM, Language Model; DNN, Deep Neural Network; FCN, Fully Convolutional Network; CNN, Convolutional Neural Network; GAP, Global Average Pooling; FC, Fully Connected; OARS, Older Americans Resources and Services; LDA, Latent Dirichlet Allocation; ADReSS, Alzheimer's Dementia Recognition through Spontaneous Speech; SVF, Semantic Verbal Fluency; NLP, Natural Language Processing; RMSE, Root-Mean-Square Error; IR, Image Recognition; GPU, Graphics Processing Unit; LSTM, Long Short-Term Memory.

TABLE 1 | Cookie Theft Picture datasets.

Dataset	Language	Total	HC	MCI	AD
ADReSS Luz et al., 2020	English	156	78		78
Pitt Corpus Becker et al., 1994	English	312	104		208
WLS Herd et al., 2014	English	1366			
IVA Mirheidari et al., 2019b	English	33	16		17
Hebrew CTP Kavé and Dassa, 2018	Hebrew	70	35		35
MECSD Wang et al., 2019a	Mandarin	85	65		20
NTU Chien et al., 2019	Mandarin	50	40		10
Swedish CTP Wallin et al., 2016	Swedish	67	36	31	
French CTP Fraser et al., 2019b	French	58	25		33

learning model is critically important to the downstream task. This customization is often based on two strategies.

- Fixed feature extractor: Remove the last one or several layers from the pre-trained model, and treat the rest of the pre-trained model as a fixed feature extractor for the downstream dataset. Then, apply a simple classification model over the features from the fixed feature extractor. The training process will only modify the weights of the classification model. The fixed feature extractor strategy can avoid the overfitting problem when the downstream dataset is small.
- Fine-tuning: Replace the last one or several layers of the pre-trained model with customized layers for the downstream task. In the training process, the weights of the pre-trained model are fine-tuned by continuing the back-propagation. In this strategy, the pre-trained model produces generic features, and the fine-tuning process modifies the model to be more specific to the details of the downstream task. The fine-tuning strategy often requires the downstream dataset to be sufficiently large to avoid the overfitting problem.

We explored transfer learning with a fine-tuning strategy for the following reasons: (i) the fine-tuning strategy relies more on the data and less on the customization of the network architecture. Specifically, for each pre-trained model, we adopted the same modification strategy, i.e., replacing the last layer with a standard fully connected (FC) layer and fine-tuning the weights of all layers with the training dataset of the downstream task. (ii) We envisioned the downstream dataset is a special task, which requires a different knowledge set from the tasks corresponding to the pre-training dataset. The fine-tuning strategy enables the training using a downstream dataset to customize the model using back-propagation, which puts more emphasis on the newly acquired knowledge. (iii) The fixed feature extractor strategies have been explored in literature (Balagopalan et al., 2020; Koo et al., 2020; Pompili et al., 2020).

Koo et al. (2020) and Pompili et al. (2020) employed transfer learning techniques to extract both acoustic and linguistic features from pre-trained models, combined these features with handcrafted features, and customized a convolutional recurrent neural network to perform the downstream tasks. Their customized network architectures, though different in detail, produced similar results and conclusions. In comparison, we did not use pre-trained models as a fixed feature extractor, but followed the fine-tuning strategy to train an end-to-end network model. Balagopalan et al. (2020) compared handcrafted features including lexico-syntactic features, acoustic features, and semantic features, with pre-trained automatic features using BERT (Devlin et al., 2018), and concluded that automatic features (83.3% accuracy) outperform the handcrafted features (75.0% accuracy). Edwards et al. (2020) explored multi-scale (word and phoneme level) audio models and their models achieved 79.2% accuracy at best, which is higher than the models using text features (i.e., Word2Vec) and multi-modal fusion. Rohanian et al. (2020) proposed a multi-modal gating mechanism to fusion audio and text features in a Long Short-Term Memory (LSTM) model and achieved a better accuracy of 79.2% compared to the LSTM model with either audio or text features (highest accuracy 73.0%). Yuan et al. (2020) explored disfluencies and fine-tuning pre-trained language models, aligned audio and text using forced alignment, and re-created the punctuation marks in the text using manually defined thresholds to identify pauses. It achieved an accuracy of 85.4% using BERT and 89.6% using ERNIE (Sun et al., 2020). We consider the thresholds used to identify pauses (Yuan et al., 2020) is still a handcrafted feature. In comparison with the above works, we avoid the complex design and evaluation of handcrafted features and the heavy network architecture. We built an end-to-end network model using the pre-trained networks and a fine-tuning strategy. In addition, Pappagari et al. (2020) employed speaker recognition and natural language processing methods. Specifically, it explored the x-vector (Snyder et al., 2018) and BERT for extracting acoustic and linguistic features, fusioned them with Gradient Boosting Regressor, and achieved 75.0% accuracy using the ADReSS training/test dataset. We considered that our selected pre-training tasks are more representative and similar to the AD classification task, compared to the speaker recognition task (Snyder et al., 2018; Pappagari et al., 2020).

In this paper, we explored a variety of transfer learning techniques and compared several transfer learning models. Note that our training and testing processes strictly followed the ADReSS challenge, i.e., we only used the ADReSS training dataset for training and reported the classification/regression results over the ADReSS testing dataset. Specifically, we investigated the following:

- **Evaluation of transfer learning**: We studied four types of pre-trained models, and customized and fine-tuned our transfer learning models based on the downstream tasks and datasets. We evaluated the impact of the similarity between the pre-training datasets and the downstream datasets on the performance.
- **Multi-modal transfer learning**: We applied a multi-modal transfer learning to incorporate inputs of both audio and text. We investigated whether the audio and text data share complementary information to further improve the performance of the downstream tasks.
- **Multi-task transfer learning**: We applied a multi-task transfer learning to output both the AD/non-AD labels and the Mini-Mental State Examination (MMSE) scores (a test assessing global cognitive functioning). We investigated

whether two downstream tasks are highly correlated and whether integrated training can reinforce the performance of the two tasks.

2. SPEECH DATASET FOR DEMENTIA RESEARCH

In the ADReSS challenge (Luz et al., 2020), a pre-processed CTP dataset from the Pitt Corpus (Becker et al., 1994) is created with the balanced groups of participants in age, gender, and cognitive status. The ADReSS training dataset includes speech data from 24 male participants with AD, 30 female with AD, 24 male non-AD participants, and 30 female non-AD participants. The ADReSS testing dataset includes speech data from 11 male participants with AD, 13 female with AD, 11 male non-AD participants, and 13 female non-AD participants. The complete dataset information can be found in Luz et al. (2020). In this paper, we studied the ADReSS dataset, i.e., we trained our models with the ADReSS training dataset and reported the performance of classification and regression tasks over the ADReSS testing dataset.

3. PRE-TRAINING DATASETS

In this section, we describe datasets in four domains, i.e., image, audio, speech, and text. These datasets have been successfully explored in their domains for enhanced performance of transfer learning models.

3.1. Image Dataset

The most commonly used large-scale image classification dataset for pre-training is ImageNet (Deng et al., 2009). ImageNet (http://image-net.org/) is an image dataset organized according to the WordNet hierarchy. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synset." There are more than 100,000 synsets in WordNet, the majority of which are nouns (80,000+). ImageNet provides, on average, 1,000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated. ImageNet pre-training has been widely used in various computer vision tasks, such as fine-grained image classification (Russakovsky et al., 2015; Fu et al., 2017; Cui et al., 2018), object detection (Redmon et al., 2016; He et al., 2017), and sense text detection (Zhou et al., 2017; Wang et al., 2019b).

3.2. Audio Dataset

AudioSet (https://research.google.com/audioset/) (Gemmeke et al., 2017) is extracted from YouTube videos. It consists of 10-s segments, and each segment is labeled by human effort. All segments are organized in 632 classes, organized in a hierarchical structure with a max depth of 6 levels. AudioSet is considered as a general audio dataset, e.g., the top-level classes include "Human sound," "Animal sounds," "Natural sounds," "Music," "Sounds of things," "Source-ambiguous sounds," and "Channel, environment and background." The dataset contains 1,789,621 segments (4,971 h) in total. AudioSet is commonly used for the pre-training of acoustic event detection (Arora and Haeb-Umbach, 2017) and sound event tagging (Diment and Virtanen, 2017).

3.3. Speech Dataset

LibriSpeech (http://www.openslr.org/12/) (Panayotov et al., 2015) is a corpus of approximately 1,000 h of 16 kHz read English speech, prepared by Vassil Panayotov with the assistance of Daniel Povey. The data are derived from reading audiobooks from the LibriVox project and has been carefully segmented and aligned. The typical usage of this dataset is for ASR (Huang et al., 2020; Zhang et al., 2020). It could also be used for self-supervised training (Chi et al., 2020; Liu et al., 2020), and transfer to the downstream task like phoneme classification, speaker recognition, and sentiment classification.

3.4. Text Dataset

BERT (https://github.com/google-research/bert) dominates NLP research by learning powerful and universal representation and utilizing self-supervised learning at the pre-training stage to encode the contextual information. The representation is beneficial to performance, especially when the data of the downstream task is limited. The pre-training datasets for BERT include the BooksCorpus (Zhu et al., 2015) (800M words) derived from textbooks and Wikipedia (2500M words) derived from Wikipedia websites. BERT (Devlin et al., 2018) and its variants (Lan et al., 2019; Liu et al., 2019; Beltagy et al., 2020) have been developed using self-supervised training for downstream tasks, e.g., text classification and question answering. Longformer (Beltagy et al., 2020) is a variant of BERT to allow the model to learn long dependencies in pre-training, and its pre-training databases additionally include one-third of a subset of the Realnews dataset (Zellers et al., 2020) with documents longer than 1,200 tokens as well as one-third of the StoryCorpus (Trinh and Le, 2018).

4. DEEP TRANSFER LEARNING MODEL

Our transfer learning models were built within three steps: (1) **pre-training**, (2) **fine-tuning**, and (3) **testing**. In the pre-training step, a model was trained with a large-sized dataset. In the fine-tuning step, we tuned the model with the ADReSS training dataset. In the testing step, we evaluated the model using the ADReSS testing dataset. In the following, we introduce the transfer learning models based on two pre-training approaches: a *supervised classification* approach and a *self-supervised learning* approach.

4.1. Supervised Classification Approach: MobileNet and YAMNet

For this approach, we explored the audio part of the ADReSS datasets. We observed the ADReSS organizers segmented the audio data into small pieces by setting the log energy threshold parameter to 65 dB with a maximum duration of 10 s from (Haider et al., 2019; Luz et al., 2020). However, there was a concern that the segmentation may cause critical time-series information loss. Any smaller speech segments hardly represent the overall speech sample. In addition, the speech continuity is removed by segmentation, making the model inaccurately capture the time-series characteristics. Thus, our approaches aimed to accommodate an entire speech sample of each participant as input and preserve the time-series characteristics of the speech, similar to works (Hershey et al., 2017; Zhang et al., 2018).

MobileNet is a lightweight network architecture that significantly reduces the computational overhead as well as parameter size by replacing the standard convolution filters with the depth-wise convolutional filters and the point-wise convolutional filters, as proposed by Howard et al. (2017). The total parameters of the MobileNet backbone are of a size 17.2 MB, significantly less than other convolutional neural networks. Considering the limited size of the speech dataset, we considered a smaller model with less complexity, such as MobileNet, which may worth being tested. MobileNet is pre-trained with the ImageNet dataset for an image classification task. The MobileNet architecture is shown at the above layer, as shown in **Figure 1**. With an RGB image as input, the output is the probability that the image belongs to each of the 1,000 classes.

MobileNet architecture: The core of MobileNet architecture is a backbone Convolutional Neural Network (CNN), which consists of a set of convolution, pooling, and activation operations. The detailed architecture can be found in the paper (Howard et al., 2017). We used the full width (1.0) MobileNet backbone pre-trained on a resolution of 128*128 images. The backbone takes an image as an input, which is 3-dimensional $(h, w, 3)$-matrix where h is height, w is width, and 3 represents the RGB channel. The backbone converts an input of $(h, w, 3)$-matrix to an output of $(h', w', 1024)$-matrix where (h', w') are functionally related to (h, w), and 1024 represents the feature channel number, i.e., the depth of the backbone CNN. The output $(h', w', 1024)$-matrix is then fed to a Global Average Pooling (GAP) layer for reducing the dimensions of h' and w' and obtaining a 1024-dimension feature. A Fully Connected (FC) layer with 1,000 neurons produces the output according to the wanted 1,000 classes. Finally, a softmax activation layer is added to produce the classification results as the probabilities for 1,000 classes that add up to 1.

Transfer learning via MobileNet: MobileNet is pre-trained for an image classification task where its input is an image, and its output is probabilities of the classes. To apply transfer learning of MobileNet to our AD classification task, in the fine-tuning and testing steps, we need to convert an audio sample to an image sample and customize the model for the AD/non-AD outputs.

1. Extracting Mel Frequency Cepstral Coefficient (MFCC) feature maps from audio samples: Mel-frequency cepstral coefficients have been widely used in speech recognition research (Muda et al., 2010). Yancheva and Rudzicz (2016) and Fraser et al. (2016) carried out an acoustic-prosodic analysis on the Pitt Corpus using 42 MFCC features. We extracted an MFCC feature map for each participant's entire speech sample. The MFCC feature map is denoted as a (p, t)-matrix where the hyper-parameter p (64) is the MFCC order, and t is related to the duration of the speech sample. We used the librosa function with a sampling rate of 22,050, a window size of 2,048, and a step size of 512. By extracting the MFCC feature maps, we converted the speech dataset to an image dataset. The advantages of MFCC feature maps include conversion from speech to MFCC feature maps can be done automatically; the silent pauses in the audio data were preserved as a distinctive feature in MFCC feature maps; and speech from the investigator and filled pauses from the participant were preserved in MFCC feature maps and shown to be important (Tóth et al., 2018). While identifying these audio segments requires expensive human efforts or customized ASR, we envision the classification model with the input of the MFCC feature maps may learn and understand the patterns of the information.

2. Customizing model for the downstream task: Our proposed model is shown at the bottom layer of **Figure 1**. Our architecture employs the pre-trained backbone CNN module from the MobileNet. Denote the MFCC feature map of the audio sample as a $(p, t, 1)$-matrix. To match with the module input, i.e., an RGB image, we duplicated the MFCC feature map twice and made the MFCC feature map as a $(p, t, 3)$-matrix. In this way, we can feed the MFCC feature map into the backbone CNN module of the MobileNet in the same way as an RGB image. The output of the backbone CNN is denoted as a $(p', t', 1024)$-matrix where (p', t') are functionally related to (p, t). We employed a GAP-2D (two-dimensional) to reduce p' dimension and t' dimension of the matrix. We then employed a fully connected layer and a softmax activation layer to produce the classification results as two probabilities for the two classes AD/non-AD that add up to 1.

Transfer learning via YAMNet: While the MobileNet architecture is pre-trained with the ImageNet dataset, Gemmeke et al. (2017) pre-trained a similar architecture using the AudioSet dataset, called YAMNet. The input of YAMNet is the Mel spectrogram from audio data with dimensions of $(p, t, 1)$. Compared to MobileNet, YAMNet might better apply to our downstream task because the pre-training dataset and the downstream dataset are both audio datasets, and the input formats to the Backbone CNN in the pre-training/fine-tuning/testing phase are kept the same, i.e., a feature vector of $(p,\ t,\ 1)$.

4.2. Self-Supervised Learning Approach: BERT

While the supervised classification approach utilizes labeled datasets, self-supervised learning approaches take advantage of unlabeled datasets for pre-training. The removal of the labeling requirement enables the model to extract knowledge from an extended range of data sources, e.g., digital books, Wikipedia, and online news. We propose a Text BERT model and a Speech BERT model for AD classification, as shown in **Figure 2**.

Transfer learning via Text BERT: BERT (Devlin et al., 2018) is a milestone in the natural language processing domain. BERT is pre-trained with BooksCorpus (Zhu et al., 2015) (800M words) and Wikipedia (2500M words). It adopts two self-supervised tasks in the **pre-training** step: Masked Language Model (MLM) and Next Sentence Prediction (NSP). Specifically, given a pair of sentences, we first put a special [CLS] token at the beginning of the first sentence and a special [SEP] token between two

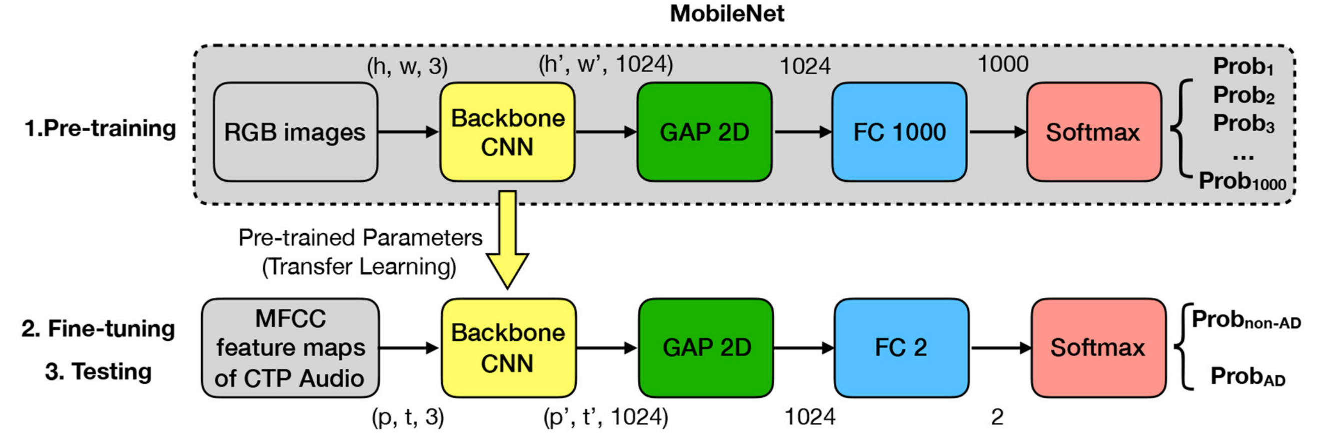


FIGURE 1 | Supervised classification approach.

sentences. Second, random masking is applied to mask a set of words with a special [MASK] token. Then the pre-processed input is fed into the BERT model, which then outputs an embedding corresponding to each input token. The pre-training is performed via the two self-supervised tasks: the MLM task aims to predict the masked words with the context; the NSP task aims to predict whether the second sentence is followed by the first sentence in the original dataset. In the **fine-tuning** and **testing** steps, the output embedding of the [CLS] token is used. To apply BERT to our AD classification task, we added a fully connected (FC) layer and a softmax activation layer to the output of the BERT model. The FC layer has two neurons, which stands for the AD and no-AD classes, respectively.

Transfer learning via Speech BERT: The Speech BERT, named Mockingjay (Liu et al., 2020), is similar to the Text BERT except for some differences: The input is the Mel spectrogram of speech data instead of the word embeddings. The pre-training task contains only the Masked Acoustic Model (MAM) task. The input does not have the [CLS] and other special tokens. Thus, instead of using output embedding of the [CLS] token for classification, we used output embeddings of all the tokens. To apply Speech BERT to our AD classification task, the output of the Speech BERT is fed into a 1D convolutional layer that convolutes through time dimension, then fed into a global average pooling layer to obtain the average through time dimension, and finally fed into an FC layer and a softmax activation layer.

5. MULTI-MODAL TRANSFER LEARNING

While Text BERT and Speech BERT models analyze text and audio datasets separately, we explored a multi-modal transfer learning via a Dual-BERT model, using both text and audio as inputs. We envision that the text and audio data of a given patient are highly related, and the outputs could reinforce each other during the training process. Dual-BERT incorporates two pre-trained BERT models, one is Text BERT and the other is the Speech BERT. As shown in **Figure 3**, the architectures of the Speech BERT and the Text BERT models remain the same as in the previous section. We further designed two types of fusion methods: Add fusion and Concat fusion. We used term "training" instead of "fine-tuning" in the following, as we mainly considered the new multi-modal transfer learning. For each fusion method, we also considered two types of training strategies, separate training and joint training.

Add fusion model: The outputs of our previous models are probabilities from the last softmax activation layer. Thus, we considered an Add fusion that adds up the outputs of the FC layers of two models, as shown in the upper part of **Figure 3**. If the Text BERT and Speech BERT models have consistent classification results, the Add fusion model outputs the result with more confidence compared to any of the two single models. On the other hand, if the two models have inconsistent classification results, the Add fusion model outputs the result that receives higher confidence from any of the two models. We considered two training strategies. (1) (Separate) We train the Text BERT and Speech BERT with text and audio, respectively. Then, the Add fusion layer will only be considered during the testing process. (2) (Joint) We train the Text BERT and Speech BERT jointly using the joint output from the Add fusion layer. The difference between these two training strategies is that the first strategy considers the confidence of the models, while the second one further considers the complementary information between text and audio data. The Add fusion part has no trainable parameters. In the separate training strategy, the training does not apply to the Add fusion part; in the joint training strategy, the Add fusion part is involved in the training process but has no parameters to be learned.

Concat fusion model: Another way to explore the multi-modal transfer learning is to concatenate the tensors of the Text BERT and Speech BERT models before the FC layer. As shown in the bottom part of **Figure 3**, after the concatenation, the Concat fusion model has an FC layer with two neurons for classification of AD/non-AD. In this model, features from text and audio are better integrated for the classification task. The Concat

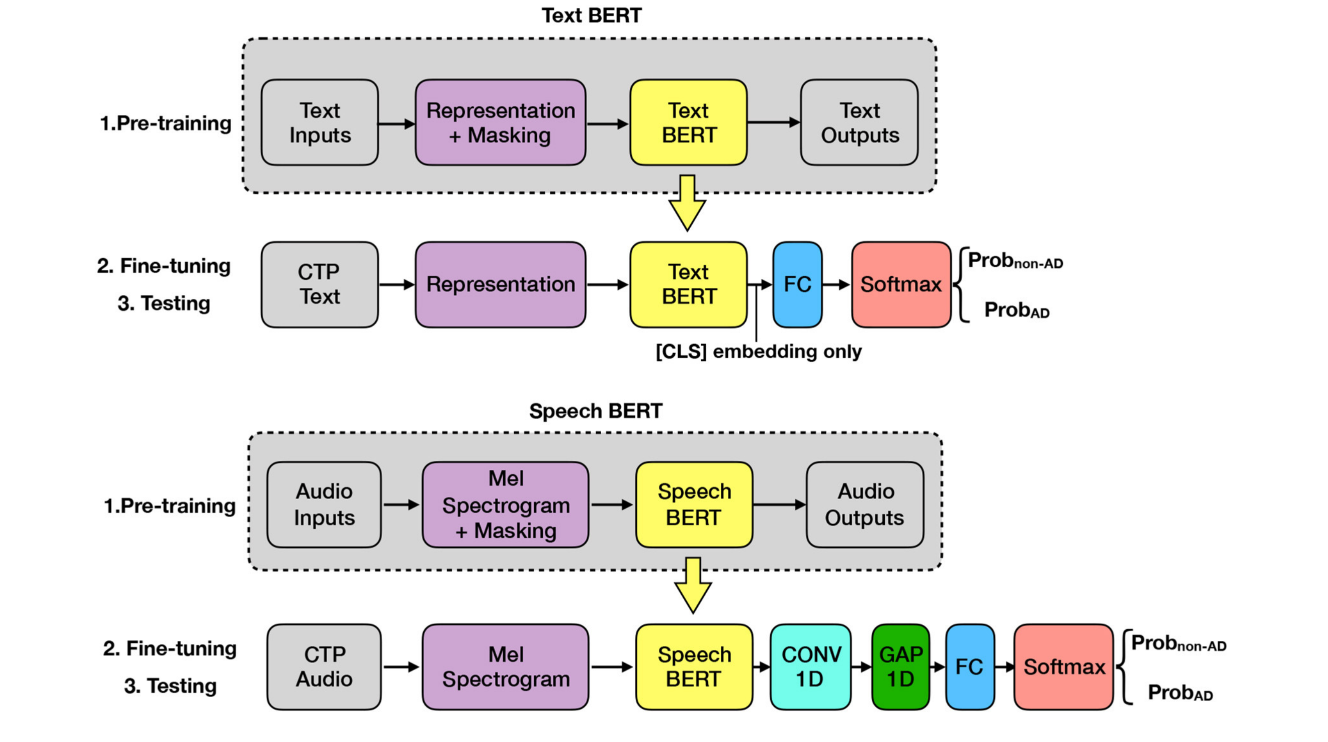


FIGURE 2 | Text BERT and Speech BERT.

fusion model always requires joint training for the additional FC layer. We have two training strategies. (1) (Separate) We train the Concat fusion model using three outputs separately. (2) (Joint) We train the Concat fusion model using the joint output only.

6. MULTI-TASK TRANSFER LEARNING

Multi-task transfer learning aims to solve multiple learning tasks at the same time while exploiting commonalities and differences across tasks. This can result in improved learning efficiency and enhanced performance for the task-specific models when compared to training the models separately.

The ADReSS challenge provides both AD/non-AD labels and MMSE scores for each data sample. In this section, we focused on the Text BERT as it produces significantly better results than the Speech BERT. As shown in the upper part of **Figure 4**, we first applied transfer learning from the Text BERT to an MMSE regression task; we placed an FC layer with a single neuron to the output of the Text BERT, and then added a Leaky ReLU layer to output the MMSE score. Since the MMSE scores are non-negative values, we adopted the Leaky Rectified Linear Unit (ReLU) activation and the mean squared error loss. The bottom part in **Figure 4** shows a multi-task transfer learning where we put an FC layer with a single neuron for the regression task and an FC layer with two neurons for the classification task. The classification task employs the softmax activation layer, and the regression task employs the Leaky ReLU activation layer. For loss functions, the classification task uses the cross-entropy loss, and the regression task uses the mean squared error loss. For training, we jointly optimized the cross-entropy loss and the mean squared error loss with the corresponding labels.

7. PERFORMANCE EVALUATION

In this section, we provide a comprehensive evaluation of the proposed deep transfer learning models. We strictly followed the ADReSS challenge (Luz et al., 2020) using the ADReSS training and testing datasets.

7.1. Implementation Details

We followed the original implementation of the pre-trained models. Specifically, the speech BERT and text BERT were implemented with PyTorch. The MobileNet and YAMNet were implemented with Tensorflow. We downloaded the pre-trained parameters of these models from online sources. For the classification task (AD/non-AD), we used the cross-entropy loss, and for the regression task (MMSE), we used the mean squared error loss. We trained our models using the Adam algorithm as optimizer (Kingma and Ba, 2014) with batch size 8 and a small learning rate of 1e-6 for models that do not use Speech BERT. For models that use Speech BERT, as our Graphics Processing Unit (GPU) resource has 32 GB memory (NVIDIA TESLA V100), we used batch size 1 to adapt our training process to the limited memory resources. We employed a fine-tuning strategy and trained all layers, including those in the pre-trained models.

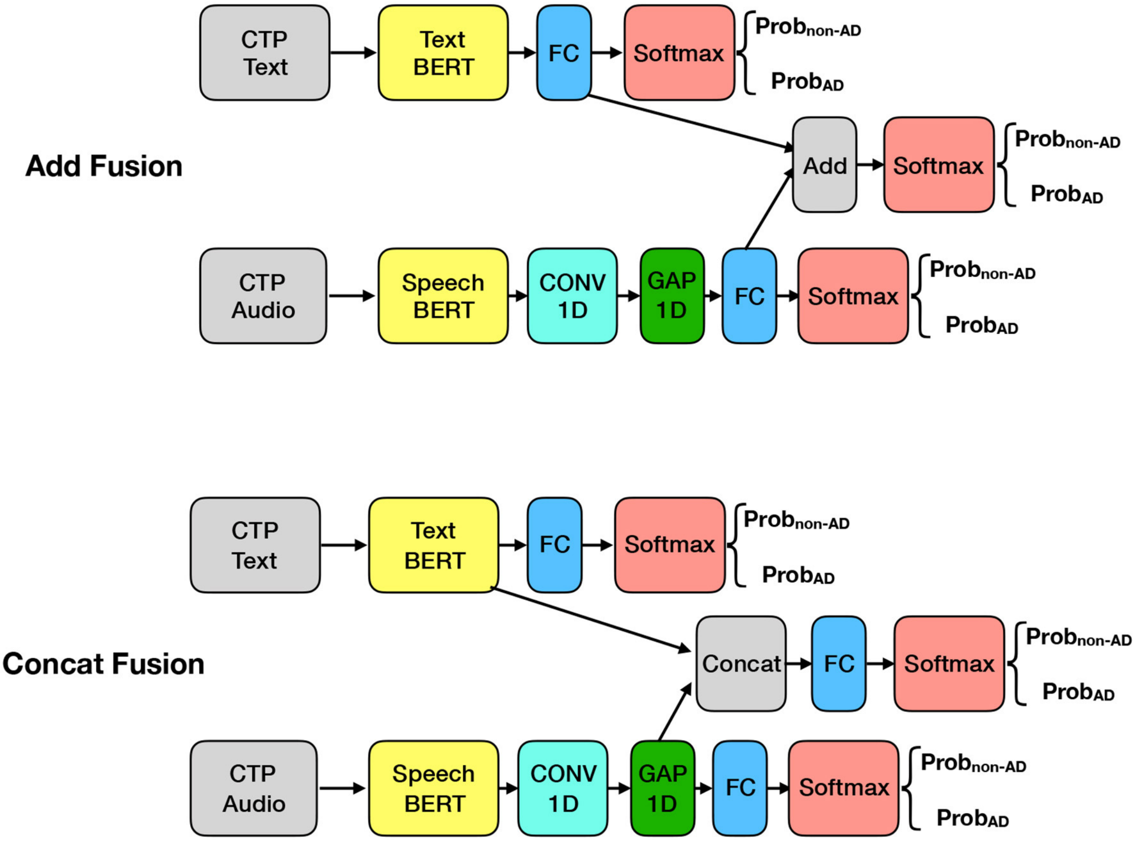


FIGURE 3 | Multi-modal transfer learning using Text/Speech BERT (Dual BERT).

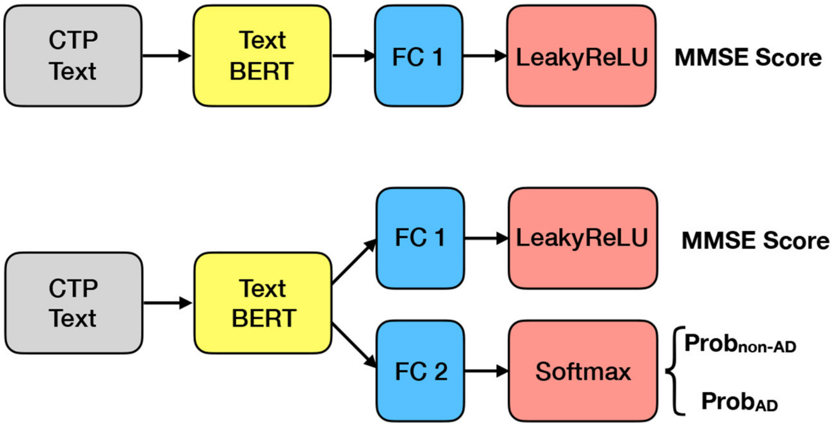


FIGURE 4 | Multi-task learning using Text BERT.

7.2. Training Strategy

Our training strategy for all models had five rounds. In each round, we used the ADReSS training dataset to train a model with a maximum of 2,000 epochs. The training stopped before reaching 2000 epochs only if the training loss was less than a pre-defined threshold of 1e-6. After the training, we selected the epoch with the smallest training loss and obtained the performance result over the ADReSS testing dataset using the selected epoch. We repeated the above process for five rounds, obtained five results, and reported their *mean and standard deviation*. We consider that the *mean and standard deviation* represent the effectiveness of the model. We also reported the best result among all epochs in five rounds to reveal the maximum potential of the models.

7.3. Evaluation Metrics

For the classification task, we employed evaluation metrics of accuracy $\frac{TN+TP}{N}$, precision $\pi = \frac{TP}{TP+FP}$, recall $\rho = \frac{TP}{TP+FN}$, and F1 score $\frac{2\pi\rho}{\pi+\rho}$, where N is the number of participants, TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively. For the regression task, we employed Root-Mean-Square Error (RMSE), the same metric used in the baseline paper provided by the ADReSS challenge.

7.4. Evaluation of Deep Transfer Learning Models

In this section, we reported the performance results of our transfer learning models with an input of audio data or text data. **MobileNet**, **YAMNet**, and **Speech BERT** were pre-trained with ImageNet, AudioSet, and LibriSpeech datasets, respectively, and were used to analyze CTP audio data. **BERT base** and **BERT large** were pre-trained with BooksCorpus, Wikipedia, and **Longformer** were pre-trained with additional Realnews and StoryCorpus. They were used to analyze CTP text data. To show the advantage of transfer learning, we also reported the performance results of the models without pre-training. The performance results are shown in **Table 2**.

MobileNet: The classification accuracy of MobileNet is 59.00 ± 5.66% without pre-training or 58.8 ± 3.49% with pre-training. Both MobileNet models achieved low accuracy, and the pre-training process surprisingly lowered the performance. We concluded the main reason is the knowledge difference between the pre-training image dataset and the CTP audio dataset. However, we found that the pre-training helped produce stable results with a lower standard deviation (from 5.66 to 3.49%). In addition, we found that Best accuracy reaches 77.08% with pre-training, much higher than 72.91% without pre-training. In other words, the model with pre-training has the potential to achieve higher accuracy, but the model cannot be fine-tuned to the optimal status due to the limited downstream dataset.

YAMNet: In general, YAMNet would be more effective than the MobileNet for our downstream task because the pre-training dataset in YAMNet is AudioSet, which is more similar to the CTP audio dataset. We confirmed this conjecture with our evaluation results of YAMNet. The classification accuracy of YAMNet without pre-training is 53.8 ± 6.88%, and the accuracy of YAMNet with pre-training is increased to 66.2 ± 4.79%. The YAMNet with pre-training resulted in a significant improvement of 12.4% compared to the same model without pre-training, which demonstrates the similarity between the AudioSet and the CTP audio dataset. In addition, the pre-training enabled the YAMNet to produce more stable outputs (from 6.88 to 4.79%) and higher Best accuracy (from 79.17 to 83.33%).

Speech BERT: Speech BERT, similar to Text BERT, employs a self-supervised learning approach. The pre-training process employs the MAM task. Speech BERT has a length restriction problem of max positional encoding in pre-training of 5,000 tokens (about 1 min). To solve this problem, in training, if the audio sample produces more than 5,000 tokens, we randomly choose a window to sample the audio for 5,000 tokens. And in the testing, we used a non-overlapped sliding window technique to sample the whole audio and averages the classification probabilities corresponding to all windows. We further filtered the audio data of the investigator to reduce the audio length, while for MobileNet/YAMNet, both audio data of the investigator and participant were kept as input.

We observed that the Speech BERT model with pre-training resulted in less accuracy 63.33%, compared to 66.67% from the model without pre-training. This finding may have been due to the Speech BERT models employing a self-supervised MAM task, which is significantly different from our downstream task (i.e., classification). Alternatively, the self-supervised MAM task aims to explore the strong correlation between the audio segments. While such a correlation in the transcript is explicit due to the language model, the correlation among audio segments might be more complicated and more challenging to be learned. In addition, the pre-training process helps to increase the potential of the model by providing a higher Best accuracy of 79.17% (> 77.08% without pre-training).

Text BERT: We considered three Text BERT models, i.e., BERT base and BERT large (Devlin et al., 2018), and Longformer (Beltagy et al., 2020). The BERT base model has 12 Transformer encoders, and the BERT large model has 24 Transformer encoders. While the BERT base and BERT large were pre-trained with a max length of 512 tokens, the Longformer were pre-trained with a max length of 4,096 tokens. Therefore, when our text sample from ADReSS datasets is converted to be larger than 512 tokens, truncation is required in the BERT base and large models. In the Longformer model, all text samples from ADReSS datasets can be encoded within 4,096 tokens, and thus truncation is not needed. In addition, the pre-training databases of Longformer additionally include longer text samples from Realnews and StoryCorpus. To adapt the ADReSS text dataset to the Text BERT models, we removed the symbols that do not appear in the pre-training dataset but appear in the ADReSS text dataset.

We found the performance results of all Text BERT models are better than the previous models on audio data. Without pre-training, BERT base achieved 76.67%. With pre-training, BERT base achieved 80.83%, BERT large achieves 81.67%, and Longformer achieves 82.08%. The corresponding Best accuracy increased from 81.25% (BERT base without pre-training) to 85.42% (BERT base), 87.50% (BERT large), and 89.58% (Longformer). These findings suggest that the Text BERT models show significantly better performance because of the similarity of the pre-training text dataset and the CTP text dataset. In addition, the Longformer resulted in improved performance because it supports the input of longer text samples without truncation and has been pre-trained with additional similar datasets.

7.5. Evaluation of Multi-Modal Transfer Learning

Focusing on evaluating multi-modal transfer learning, we expected the joint training using both audio data and text data to improve the performance results of previous models. In **Table 3**,

TABLE 2 | AD Classification results using audio or text and with or without pre-training.

Model	Pre-training dataset	Classes	Precision %	Recall %	F1 %	Accuracy %	Best %
Audio Luz et al., 2020	–	non-AD	67	50	57	62	–
		AD	60	75	67		
MobileNet	–	non-AD	60.40 ± 7.86	58.40 ± 22.76	56.20 ± 13.79	59.00 ± 5.66	72.91
		AD	61.40 ± 6.89	59.80 ± 21.07	57.60 ± 10.54		
	ImageNet	non-AD	72.80 ± 6.97	28.00 ± 8.15	40.40 ± 9.85	58.80 ± 3.49	77.08
		AD	55.80 ± 2.48	90.40 ± 1.96	69.00 ± 1.67		
YAMNet	–	non-AD	52.20 ± 11.74	19.80 ± 22.61	24.60 ± 22.81	53.80 ± 6.88	79.17
		AD	53.40 ± 5.95	87.60 ± 9.56	65.80 ± 1.33		
	AudioSet	non-AD	69.60 ± 6.80	59.20 ± 7.73	63.40 ± 5.57	66.20 ± 4.79	83.33
		AD	64.40 ± 3.93	73.40 ± 8.82	68.60 ± 4.84		
Speech BERT	–	non-AD	67.74 ± 3.69	64.17 ± 3.34	65.82 ± 2.68	66.67 ± 2.95	77.08
		AD	65.84 ± 2.43	69.16 ± 5.65	67.39 ± 3.71		
	LibriSpeech	non-AD	66.13 ± 4.12	55.00 ± 4.86	59.94 ± 3.78	63.33 ± 3.12	79.17
		AD	61.48 ± 2.76	71.67 ± 4.86	66.12 ± 3.08		
Text Luz et al., 2020	–	non-AD	70	87	78	75	–
		AD	83	62	71		
BERT base	–	non-AD	78.12 ± 1.98	74.17 ± 3.12	76.05 ± 1.82	76.67 ± 1.56	81.25
		AD	75.47 ± 2.08	79.17 ± 2.63	77.23 ± 1.50		
	BooksCorpus/Wiki	non-AD	78.46 ± 1.89	85.00 ± 2.04	81.60 ± 1.96	80.83 ± 2.04	85.42
		AD	83.64 ± 2.22	76.67 ± 2.04	80.00 ± 2.13		
BERT large	BooksCorpus/Wiki	non-AD	83.05 ± 5.12	80.00 ± 3.12	81.40 ± 3.09	81.67 ± 3.34	87.50
		AD	80.65 ± 2.66	83.33 ± 5.89	81.89 ± 3.64		
Longformer	BooksCorpus/Wiki/ Realnews/Stories	non-AD	77.87 ± 3.75	90.00 ± 2.04	83.44 ± 2.33	82.08 ± 2.83	89.58
		AD	88.14 ± 2.09	74.17 ± 5.53	80.44 ± 3.55		

AD, Alzheimer's disease. Accuracy: mean and standard deviation of results of 5 rounds. Best: highest accuracy of all epochs in 5 rounds.

TABLE 3 | AD Classification results of multi-modal learning using both audio and text.

Model	Fusion / Training	Classes	Precision %	Recall %	F1 %	Accuracy %	Best %
Speech BERT	–	non-AD	66.13 ± 4.12	55.00 ± 4.86	59.94 ± 3.78	63.33 ± 3.12	79.17
		AD	61.48 ± 2.76	71.67 ± 4.86	66.12 ± 3.08		
BERT base	–	non-AD	78.46 ± 1.89	85.00 ± 2.04	81.60 ± 1.96	80.83 ± 2.04	85.42
		AD	83.64 ± 2.22	76.67 ± 2.04	80.00 ± 2.13		
Dual BERT	Add/Joint	non-AD	78.63 ± 1.77	85.83 ± 2.04	82.07 ± 1.79	81.25 ± 1.86	85.42
		AD	84.41 ± 2.13	76.67 ± 2.04	80.35 ± 1.95		
	Add/Separate	non-AD	78.96 ± 1.57	87.50 ± 2.64	82.99 ± 1.68	82.08 ± 1.66	85.42
		AD	86.05 ± 2.60	76.67 ± 2.04	81.06 ± 1.69		
	Concat/Separate	non-AD	80.39 ± 1.56	85.00 ± 3.33	82.57 ± 1.26	82.08 ± 1.02	85.42
		AD	84.21 ± 2.52	79.17 ± 2.63	81.54 ± 1.01		
	Concat/ Joint	non-AD	80.36 ± 2.06	85.00 ± 2.04	82.59 ± 1.56	82.08 ± 1.66	87.50
	(No pre-train speech)	AD	84.10 ± 1.91	79.17 ± 2.63	81.53 ± 1.83		
	Concat/Joint	non-AD	78.83 ± 4.18	88.33 ± 4.09	83.15 ± 1.79	82.08 ± 2.12	89.58
	(Longformer)	AD	86.95 ± 3.38	75.83 ± 6.12	80.79 ± 2.74		
	Concat/Joint	non-AD	80.02 ± 1.16	86.67 ± 1.67	83.20 ± 1.01	82.50 ± 1.02	85.42
		AD	85.48 ± 1.46	78.34 ± 1.67	81.74 ± 1.10		
	Concat/Joint	non-AD	83.62 ± 4.25	82.50 ± 5.53	82.80 ± 1.76	82.92 ± 1.56	87.50
	(BERT large)	AD	83.04 ± 3.97	83.33 ± 5.89	82.92 ± 1.86		
YAMNet +	Concat/Joint	non-AD	78.06 ± 2.53	85.83 ± 2.04	81.76 ± 2.22	80.83 ± 2.43	89.58
BERT base		AD	82.70 ± 3.65	82.50 ± 5.53	82.45 ± 3.07		

AD, Alzheimer's disease. Accuracy: mean and standard deviation of results of 5 rounds. Best: highest accuracy of all epochs in 5 rounds.

we list the performance results of 10 models. The first one is BERT base, and the second one is Speech BERT, which was evaluated in the previous section. Their performance results will serve as a baseline. The next seven models are variants of the Dual BERT models. Their architectures are a combination of a Speech BERT model and a Text BERT model. As discussed in section 5, Dual BERT can employ the Add fusion or the Concat fusion to combine the Speech BERT and the Text BERT, and can be trained with a separate training strategy or a joint training strategy. The last multi-modal transfer learning replaced Speech BERT with YAMNet as YAMNet achieves an accuracy (66.2%) higher than Speech BERT (63.33%).

The following observations were made:

- All seven Dual BERT models achieved higher classification accuracy than the two baseline models, confirming that the text data and audio data have complementary information that can be jointly learned by the model for improved performance.
- Concat fusion achieved higher classification accuracy than Add fusion. While the Add fusion picks one model with higher confidence in the classification results, the Concat fusion aims to merge the features of both text data and audio data for a hybrid representation. The performance gain of the Concat fusion further confirms the complementary information between the text data and audio data.
- From the previous analysis, we found the Speech BERT without pre-training achieved a higher accuracy (66.67%) than the Speech BERT with pre-training (63.33%). Thus, we evaluate a multi-modal transfer learning model using the Speech BERT without pre-training and BERT base with pre-training. As shown in **Table 3**, we confirm that the pre-training of Speech BERT helps the multi-modal transfer learning to achieve a higher accuracy (82.50%), compared to the Dual BERT without pre-training on speech model (82.08%).
- From the previous analysis, we found BERT large (81.67%) and Longformer (82.08%) outperformed BERT base (80.83%). Thus, we replaced BERT base with BERT large and Longformer in the Dual BERT. While the multi-modal transfer learning using BERT large achieved the highest accuracy (82.92%), the multi-modal transfer learning using Longformer achieves the highest Best accuracy (89.58%).
- From the previous analysis, we found the YAMNet yielded the highest accuracy result (66.20%) among all the models using audio data. Thus, we evaluated a multi-modal transfer learning using the YAMNet and BERT base. However, this model did not outperform any of the Dual BERT models.

7.6. Evaluation of Multi-Task Transfer Learning

Relation between MMSE regression and AD classification: Given the ADReSS dataset, we explored a threshold-based strategy to understand the relation between the MMSE scores and AD/non-AD labels. We set a threshold T on MMSE scores to infer AD/non-AD status. If a patient's MMSE score is less than T, the patient's data are labeled with AD; if a patient's MMSE score is larger or equal to T, the patient's data are labeled with non-AD. We reported the performance result of the threshold-based strategy over the ADReSS training/testing dataset separately in **Figure 5** and **Table 4**. We found that for the ADReSS training dataset, the highest accuracy is 97.2% at a threshold of 26, and for the ADReSS testing dataset, the highest accuracy is 91.67% at a threshold of 28. If we adopt the threshold of 26 from the training dataset and apply it to the testing dataset, the threshold-based strategy results in an accuracy of 89.58%, which is the upper bound that multi-task transfer learning theoretically can achieve. According to the CTP dataset description (Becker et al., 1994), the patients with AD have an MMSE score in the range of 8–30, while the patients with non-AD have an MMSE score in the range of 26–30. The AD labels are determined from seven cognitive domains, including memory, construction, perception, attention, language, orientation, and executive functions. In comparison, the MMSE is a 30-point widely used cognitive screening measure, taking about 10 min to administer. In our evaluation, given the limited number of data samples, a small number of inconsistent cases might produce a negative impact on

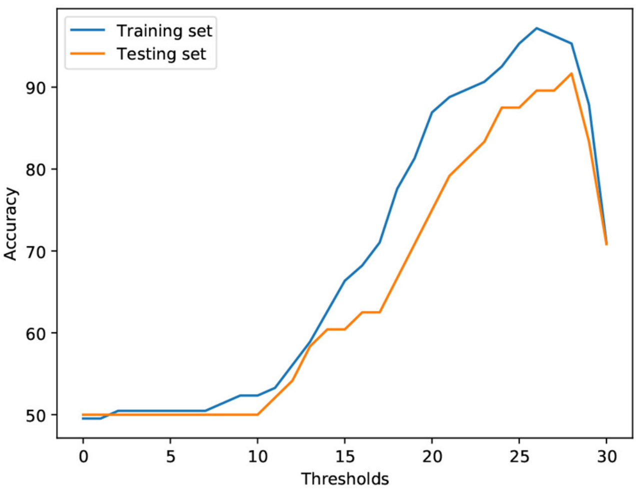


FIGURE 5 | Threshold-based strategy (0–30).

TABLE 4 | Threshold-based strategy (20–30).

T	Accuracy (Training)	Accuracy (Testing) %
20	86.92	75.00
21	88.79	79.17
22	89.72	81.25
23	90.65	83.33
24	92.52	87.50
25	95.33	87.50
26	**97.20**	**89.58**
27	96.26	89.58
28	95.33	**91.67**
29	87.85	83.33
30	71.03	70.83

The highest accuracy in training, the highest accuracy in testing, and the testing accuracy corresponding to the highest accuracy in training are in bold.

TABLE 5 | Classification and regression results of multi-task transfer learning using CTP text.

Model	Pre-training	Settings	Accuracy %	Best %	RMSE	Best RMSE
Text Luz et al., 2020	–	Classification	75	–		
	–	Regression			5.20	–
BERT base	No	Classification	76.67 ± 1.56	81.25	–	–
		Regression	–	–	5.18 ± 0.04	4.65
		Multi-task	78.75 ± 1.56	83.33	4.70 ± 0.02	4.39
	Yes	Classification	80.83 ± 2.04	85.42	–	–
		Regression	–	–	4.15 ± 0.01	4.06
		Multi-task	80.83 ± 1.56	87.50	4.96 ± 0.01	4.20

AD, Alzheimer's disease. Accuracy: mean and standard deviation of results of 5 rounds. Best: highest accuracy of all epochs in 5 rounds. RMSE: mean and standard deviation of root-mean-square errors of 5 rounds. Best RMSE: lowest RMSE of all epochs in 5 rounds.

the joint training process when the outputs of single-task models are highly consistent with the corresponding labels/scores.

We focused on evaluating the proposed multi-task transfer learning, which is built on the BERT base model with an input of the CTP text data. One challenge of the multi-task transfer learning model is the imbalanced loss from the AD classification task and the MMSE regression task. Denote the regression loss (mean squared error) as l_{mse} and the classification loss (cross-entropy) as l_{ce}. We define the total loss of the multi-task transfer learning model as $l = \lambda l_{mse} + l_{ce}$, where λ is a balance factor to avoid the unbalanced impact between the classification loss and regression loss. In our experiment, we set $\lambda = 0.01$.

We evaluated a regression model and a multi-task transfer learning model using BERT base. As shown in **Table 5**, when using BERT base without pre-training, the multi-task transfer learning model outperformed the single-task models, i.e., the classification accuracy is increased from 76.67 to 78.75%, and the RMSE decreased from 5.18 to 4.70. The evaluation results confirmed that the two tasks help each other to achieve a better performance, especially when both single-task models have room to be improved. In comparison, when using BERT base with pre-training, the multi-task transfer learning model introduced limited performance gain in classification and introduced a negative impact in the regression model. Specifically, the average classification accuracy remained the same at 80.83%, the standard deviation decreased from 2.04 to 1.56%, and Best accuracy is increased from 85.42 to 87.50%, close to the accuracy of 89.58% from the threshold-based strategy. For classification, multi-task learning kept the training more stable and increased the maximal potential of the model, and the MMSE scores provide a limited positive impact on the AD classification task. For regression, RMSE increased from 4.15 to 4.96, which reveals a negative impact of the joint training. This may have been due to the inconsistent cases of MMSE scores and AD/non-AD labels, and the MMSE regression task is more fined-grained and thus received a stronger impact from the inconsistent cases.

7.7. Summary of Best Cases Using Transfer Learning

Table 6 shows the best cases of our experiments of text-based, audio-based, and multi-modal transfer learning models. The best case of the audio model achieved 66.20%, while the best case of the text model achieved 82.08%. We consider that the performance gain of the text model may be due to the high similarity between the pre-training text dataset and the CTP text dataset. In addition, the multi-modal model using both audio and text achieved the highest accuracy of 82.92% in its best case, demonstrating that audio and text data provided complementary information. Our multi-task model achieved an accuracy of 80.83%, lower than the accuracy of the text-based model and the multi-modal model. We consider that the performance degradation of the multi-task model may be due to the inconsistency between labels and scores that were used in multiple tasks.

8. CONCLUSIONS

We explored transfer learning techniques for an AD classification task and an MMSE regression task. The transfer learning models were pre-trained with general large-sized datasets, and fine-tuned and tested using the ADReSS datasets. Our models had minimal customization and mostly relied on the training data and fine-tuning process to incorporate the knowledge of the downstream task into the pre-trained model. From our comprehensive evaluation, we drew the following three conclusions.

8.1. Transfer Learning on Text Data Results in High Accuracy, but Transfer Learning on Audio Data Might Have More Potential

Our findings showed that the transfer learning on text data achieved high accuracy in the downstream tasks and always outperformed the transfer learning on audio data. This suggests that the transfer learning model understands the text better than the audio. We considered the text data are generated from the audio data through human transcribing effort. Thus, the additional information that the text data contain, but not the audio data contain, might be the transcriber's knowledge in the transcribing process. The transcriber extracts task-specific information, such as the CTP and information units in the photo. However, while the text data implicitly contain the transcriber's knowledge, the audio data do not contain. And our training process of the transfer learning models on audio data does not take advantage of the transcriber's knowledge. We expect that the task-specific information is highly useful, and our transfer learning models on audio data can be further improved by integrating such information. In addition, different parts of the text might be highly relevant, but the relevance of different audio segments might be unclear and difficult to

TABLE 6 | The best classification cases of the audio-based, text-based, and multi-modal models.

Input	Model (with pre-training)	Classes	Precision %	Recall %	F1 %	Accuracy %	Best %
Audio	YAMNet	non-AD	69.60 ± 6.80	59.20 ± 7.73	63.40 ± 5.57	66.20 ± 4.79	83.33
		AD	64.40 ± 3.93	73.40 ± 8.82	68.60 ± 4.84		
Text	Longformer	non-AD	77.87 ± 3.75	90.00 ± 2.04	83.44 ± 2.33	82.08 ± 2.83	89.58
		AD	88.14 ± 2.09	74.17 ± 5.53	80.44 ± 3.55		
Audio + Text	Dual BERT Concat / Joint (BERT large)	non-AD	83.62 ± 4.25	82.50 ± 5.53	82.80 ± 1.76	82.92 ± 1.56	87.50
		AD	83.04 ± 3.97	83.33 ± 5.89	82.92 ± 1.86		

AD: Alzheimer's disease. Accuracy: mean and standard deviation of results of 5 rounds. Best: highest accuracy of all epochs in 5 rounds.

be learned by the proposed models. Thus, we concluded that the low accuracy of the transfer learning on audio data was likely observed because the introduced pre-trained models did not extract good representation from the audio data from the downstream perspective. However, we envision that the future large-sized speech datasets might contain audio data and auto-translated text data via ASR. For example, the larger CTP dataset WLS (Herd et al., 2014) contains text data from Kaldi ASR. Thus, our future work on transfer learning aims to explore a better pre-trained model, including supervised ASR models and self-supervised audio models.

8.2. Multi-Modal Transfer Learning Reveals Complementary Information of Text and Audio

Our multi-modal transfer learning introduced a slight but not significant improvement in terms of accuracy, demonstrating that the audio and text data provide complementary information. Specifically, while the text model alone already achieved high accuracy, adding the analysis of audio data can improve performance results almost in every case. More importantly, if we consider that the text data contain semantic information only, the complementary information that the audio data contain, but not the text data contain, might be the non-semantic information, such as filled pause, silent pause, and other implicit features. The non-semantic information may or may not be used to implement effective classification alone, but they should be useful if they are jointly analyzed with the semantic information. We envision that the model can be improved if it learns the positional information of both semantic and non-semantic features, e.g., the pause information between words or between sentences.

8.3. Multi-Task Transfer Learning Reveals Positive and Negative Impacts on AD Classification and MMSE Regression

Our multi-task transfer learning of the classification and regression tasks yielded significantly better performance when both single-task models did not perform well. The performance gain is obtained due to the consistency between most MMSE scores and the AD/non-AD labels. However, when the outputs of the single-task models are highly consistent with the corresponding labels/scores, the performance of multi-task learning declined due to a small number of samples with inconsistent scores and labels. This suggests the need to investigate the meaning behind the AD classification task and the MMSE regression task. The AD/non-AD labels seem coarse-grained, but they are generated by evaluating patients on several cognitive domains. The MMSE is less accurate and considered a screening measure of global cognitive functioning. We confirmed that such inconsistency existed by exploring a threshold-based strategy on the ADReSS training and testing datasets. Thus, we considered that multi-task transfer learning produces a limited impact on accuracy improvement due to the inconsistency between labels and scores. In conclusion, we believe that the deep transfer learning techniques need to be simple, comparable, and applicable to newer tasks, larger datasets, and heterogeneous labels to produce a long-lasting impact in dementia research.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://www.homepages.ed.ac.uk/sluzfil/ADReSS/.

AUTHOR CONTRIBUTIONS

YZ and XL: technique design, experiments, evaluation, and paper writing. JB and RR: dementia expert knowledge, evaluation, and paper writing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the US National Institutes of Health National Institute on Aging, under grant no. R01AG067416.

REFERENCES

Alzheimer's Association (2020). *2021 Alzheimer's Disease Facts And Figures. Special Report: Race, Ethnicity And Alzheimer's In America*. Available online at: https://www.alz.org/media/Documents/alzheimers-facts-and-figures.pdf

Arora, P., and Haeb-Umbach, R. (2017). "A study on transfer learning for acoustic event detection in a real life scenario," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)* (Luton: IEEE), 1–6. doi: 10.1109/MMSP.2017.8122258

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection. *arXiv [Preprint].* arXiv:2008.01551. Available online at: https://arxiv.org/abs/1909.11942 doi: 10.21437/Interspeech.2020-2557

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol.* 51, 585–594. doi: 10.1001/archneur.1994.00540180063015

Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V., and Croteau, J. (2017). Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimer's type dementia in older adults: a systematic review and meta-analysis. *Neuropsychol. Rev.* 27, 328–353. doi: 10.1007/s11065-017-9361-5

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: the long-document transformer. *arXiv [Preprint].* arXiv:2004.05150. Available online at: https://arxiv.org/abs/2004.05150

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., and Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: a review. *Front. Psychol.* 8:269. doi: 10.3389/fpsyg.2017.00269

Chi, P.-H., Chung, P.-H., Wu, T.-H., Hsieh, C.-C., Li, S.-W., and Lee, H.-Y. (2020). Audio albert: a lite bert for self-supervised learning of audio representation. *arXiv [Preprint].* arXiv:2005.08575. doi: 10.1109/SLT48900.2021.9383575

Chien, Y.-W., Hong, S.-Y., Cheah, W.-T., Yao, L.-H., Chang, Y.-L., and Fu, L.-C. (2019). An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network. *Sci. Rep.* 9, 1–10. doi: 10.1038/s41598-019-56020-x

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., et al. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain Lang.* 53, 1–19. doi: 10.1006/brln.1996.0033

Cui, Y., Song, Y., Sun, C., Howard, A., and Belongie, S. (2018). "Large scale fine-grained categorization and domain-specific transfer learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE). 4109–4118. doi: 10.1109/CVPR.2018.00432

de la Fuente Garcia, S., Ritchie, C. W., and Luz, S. (2019). Protocol for a conversation-based analysis study: prevent-ed investigates dialogue features that may help predict dementia onset in later life. *BMJ Open* 9:e026254. doi: 10.1136/bmjopen-2018-026254

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv [Preprint].* arXiv:1810.04805. Available online at: https://arxiv.org/abs/1810.04805

Diment, A., and Virtanen, T. (2017). "Transfer learning of weakly labelled audio," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (New Paltz, NY: IEEE), 6–10. doi: 10.1109/WASPAA.2017.8169984

Edwards, E., Dognin, C., Bollepalli, B., and Singh, M. (2020). "Multiscale system for Alzheimer's dementia recognition through spontaneous speech," in *Interspeech 2020 (ISCA)* (Shanghai), 2197–2201. doi: 10.21437/Interspeech.2020-2781

Fraser, K., Rudzicz, F., Graham, N., and Rochon, E. (2013). "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies* (Grenoble: Association for Computational Linguistics), 47–54. Available online at: https://www.aclweb.org/anthology/W13-3909 (accessed April 22, 2021).

Fraser, K. C., Fors, K. L., and Kokkinakis, D. (2019a). Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Comput. Speech Lang.* 53, 121–139. doi: 10.1016/j.csl.2018.07.005

Fraser, K. C., Linz, N., Li, B., Lundholm Fors, K., Rudzicz, F., König, A., et al. (2019). "Multilingual prediction of Alzheimer's disease through domain adaptation and concept-based language modelling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers) (Minneapolis, MN: Association for Computational Linguistics), 3659–3670. doi: 10.18653/v1/N19-1367

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49, 407–422. doi: 10.3233/JAD-150520

Fu, J., Zheng, H., and Mei, T. (2017). "Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE), 4476–4484. doi: 10.1109/CVPR.2017.476

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., et al. (2017). "Audio set: an ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA: IEEE), 776–780. doi: 10.1109/ICASSP.2017.7952261

Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., et al. (2016). "Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection," in *Interspeech* (San Francisco, CA), 107–111. doi: 10.21437/Interspeech.2016-384

Haider, F., De La Fuente, S., and Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE J. Sel. Top. Signal Process.* 14, 272–281. doi: 10.1109/JSTSP.2019.2955022

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Venice), 2961–2969. Available at: https://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html (accessed April 22, 2021).

Herd, P., Carr, D., and Roan, C. (2014). Cohort profile: Wisconsin longitudinal study (wls). *Int. J. Epidemiol.* 43, 34–41. doi: 10.1093/ije/dys194

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA: IEEE), 131–135. doi: 10.1109/ICASSP.2017.7952132

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., and Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *Int. J. Speech Lang. Pathol.* 12, 29–34. doi: 10.3109/17549500903137256

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv [Preprint].* arXiv:1704.04861. Available online at: https://arxiv.org/abs/1704.04861

Huang, W., Hu, W., Yeung, Y. T., and Chen, X. (2020). Conv-transformer transducer: low latency, low frame rate, streamable end-to-end speech recognition. *arXiv [Preprint].* arXiv:2008.05750. doi: 10.21437/Interspeech.2020-2361

Kavé, G., and Dassa, A. (2018). Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology* 32, 27–40. doi: 10.1080/02687038.2017.1303441

Kim, B. S., Kim, Y. B., and Kim, H. (2019). Discourse measures to differentiate between mild cognitive impairment and healthy aging. *Front. Aging Neurosci.* 11:221. doi: 10.3389/fnagi.2019.00221

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint].* arXiv:1412.6980. Available online at: https://arxiv.org/abs/1412.6980

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition. *arXiv [Preprint].* arXiv:2009.04070. Available online at: https://arxiv.org/abs/2009.04070

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: a lite bert for self-supervised learning of language representations. *arXiv [Preprint].* arXiv:1909.11942.

Liu, A. T., Yang, S., Chi, P.-H., Hsu, P., and Lee, H. (2020). "Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 6419–6423. doi: 10.1109/ICASSP40776.2020.9054458

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: a robustly optimized bert pretraining approach. *arXiv [Preprint].* arXiv:1907.11692. Available online at: https://arxiv.org/abs/1907.11692

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). "Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge," in *Interspeech 2020 (ISCA)*, 2172–2176. doi: 10.21437/Interspeech.2020-2571

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., et al. (2017). "An avatar-based system for identifying individuals likely to develop dementia," in *Interspeech 2017 (ISCA)* (Stockholm), 3147–3151. doi: 10.21437/Interspeech.2017-690

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019a). Dementia detection using automatic analysis of conversations. *Comput. Speech Lang.* 53, 65–79. doi: 10.1016/j.csl.2018.07.006

Mirheidari, B., Pan, Y., Walker, T., Reuber, M., Venneri, A., Blackburn, D., et al. (2019b). Detecting Alzheimer's disease by estimating attention and elicitation path through the alignment of spoken picture descriptions with the picture prompt. *arXiv [Preprint]*. arXiv:1910.00515. Available online at: https://arxiv.org/abs/1910.00515

Muda, L., Begam, M., and Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv [Preprint]*. arXiv:1003.4083. Available online at: https://arxiv.org/abs/1003.4083

Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018a). Connected speech and language in mild cognitive impairment and alzheimer's disease: a review of picture description tasks. *J. Clin. Exp. Neuropsychol.* 40, 917–939. doi: 10.1080/13803395.2018.1446513

Mueller, K. D., Koscik, R. L., Hermann, B. P., Johnson, S. C., and Turkstra, L. S. (2018b). Declines in connected language are associated with very early mild cognitive impairment: results from the wisconsin registry for Alzheimer's prevention. *Front. Aging Neurosci.* 9:437. doi: 10.3389/fnagi.2017.00437

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (South Brisbane, QLD: IEEE), 5206–5210. doi: 10.1109/ICASSP.2015.7178964

Pappagari, R., Cho, J., Moro-Velazquez, L., and Dehak, N. (2020). "Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity," in *Interspeech 2020 (ISCA)* (Shanghai), 2177–2181. doi: 10.21437/Interspeech.2020-2587

Pompili, A., Rolland, T., and Abad, A. (2020). The inesc-id multi-modal system for the address 2020 challenge. *arXiv [Preprint]*. arXiv:2005.14646. doi: 10.21437/Interspeech.2020-2833

Prado, C. E., Watt, S., Treeby, M. S., and Crowe, S. F. (2019). Performance on neuropsychological assessment and progression to dementia: a meta-analysis. *Psychol. Aging* 34:954. doi: 10.1037/pag0000410

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 779–788. Available online at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html (accessed April 22, 2021).

Rohanian, M., Hough, J., and Purver, M. (2020). "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," in *Interspeech 2020 (ISCA)* (Shanghai), 2187–2191. doi: 10.21437/Interspeech.2020-2721

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Sajjadi, S. A., Patterson, K., Tomek, M., and Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology* 26, 847–866. doi: 10.1080/02687038.2012.654933

Sattler, C., Wahl, H.-W., Schrder, J., Kruse, A., Schönknecht, P., Kunzmann, U., et al. (2015). "Interdisciplinary longitudinal study on adult development and aging (ILSE)," in *Encyclopedia of Geropsychology*, ed N. A. Pachana (Singapore: Springer), 1–10. doi: 10.1007/978-981-287-080-3_238-1

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). "X-vectors: robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 5329–5333. doi: 10.1109/ICASSP.2018.8461375

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., et al. (2020). "ERNIE 2.0: a continual pre-training framework for language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 (New York, NY: AAAI), 8968–8975. doi: 10.1609/aaai.v34i05.6428

Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2020). Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *PLos ONE* 15:e0236009. doi: 10.1371/journal.pone.0236009

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., and Szatlóczki, G. (2015). "Automatic detection of mild cognitive impairment from spontaneous speech using ASR," in *Interspeech 2015* (Drezda: ISCA), 2694–2698. Available online at: http://real.mtak.hu/27647/ (accessed April 22, 2021).

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* 15, 130–138. doi: 10.2174/1567205014666171121114930

Trinh, T. H., and Le, Q. V. (2018). A simple method for commonsense reasoning. *arXiv [Preprint]*. arXiv:1806.02847. Available online at: https://arxiv.org/abs/1806.02847

Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., et al. (2016). The gothenburg mci study: design and distribution of alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *J. Cereb. Blood Flow & Metab.* 36, 114–131. doi: 10.1038/jcbfm.2015.147

Wang, T., Yan, Q., Pan, J., Zhu, F., Su, R., Guo, Y., et al. (2019a). "Towards the speech features of early-stage dementia: design and application of the mandarin elderly cognitive speech database," in *Interspeech 2019* (Graz: ISCA), 4529–4533. doi: 10.21437/Interspeech.2019-2453

Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., et al. (2019b). "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 8439–8448. doi: 10.1109/ICCV.2019.00853

Yancheva, M., and Rudzicz, F. (2016). "Vector-space topic models for detecting Alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (Berlin: Association for Computational Linguistics), 2337–2346. doi: 10.18653/v1/P16-1221

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease". in *Interspeech 2020 (ISCA)* (Shanghai), 2162–2166. doi: 10.21437/Interspeech.2020-2516

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., et al. (2020). Defending against neural fake news. *arXiv [Preprint]*. arXiv:1905.12616.

Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., et al. (2020). "Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 7829–7833. doi: 10.1109/ICASSP40776.2020.9053896

Zhang, Y., Du, J., Wang, Z., Zhang, J., and Tu, Y. (2018). "Attention Based Fully Convolutional Network for Speech Emotion Recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (Honolulu, HI: IEEE), 1771–1775. doi: 10.23919/APSIPA.2018.8659587

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., et al. (2017). "EAST: an efficient and accurate scene text detector," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE), 2642–2651. doi: 10.1109/CVPR.2017.283

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., et al. (2015). "Aligning books and movies: towards story-like visual explanations by watching movies and reading books," in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago: IEEE), 19–27. doi: 10.1109/ICCV.2015.11

Conflict of Interest: The 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.

frontiers
in Aging Neuroscience

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fnagi.2021.642033

Language Impairment in Alzheimer's Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning

Hali Lindsay[1*†], *Johannes Tröger*[1,2†] *and Alexandra König*[3,4]

[1]German Research Center for Artificial Intelligence, DFKI GmbH, Saarbrücken, Germany, [2]ki elements, Saarbrücken, Germany, [3]Institut national de recherche en informatique et en automatique (INRIA), Stars Team, Sophia Antipolis, Valbonne, France, [4]CoBteK (Cognition-Behavior-Technology) Lab, FRIS—University Côte d'azur, Nice, France

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Oliver Peters,
Charité–Universitätsmedizin Berlin,
Germany
Serguei Pakhomov,
University of Minnesota Twin Cities,
United States

***Correspondence:**
Hali Lindsay
hali.lindsay@dfki.de

[†]These authors have contributed equally to this work and share first authorship

Received: *15 December 2020*
Accepted: *12 April 2021*
Published: *19 May 2021*

Citation:
Lindsay H, Tröger J and König A (2021) Language Impairment in Alzheimer's Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning.
Front. Aging Neurosci. 13:642033.
doi: 10.3389/fnagi.2021.642033

Alzheimer's disease (AD) is a pervasive neurodegenerative disease that affects millions worldwide and is most prominently associated with broad cognitive decline, including language impairment. Picture description tasks are routinely used to monitor language impairment in AD. Due to the high amount of manual resources needed for an in-depth analysis of thereby-produced spontaneous speech, advanced natural language processing (NLP) combined with machine learning (ML) represents a promising opportunity. In this applied research field though, NLP and ML methodology do not necessarily ensure robust clinically actionable insights into cognitive language impairment in AD and additional precautions must be taken to ensure clinical-validity and generalizability of results. In this study, we add generalizability through multilingual feature statistics to computational approaches for the detection of language impairment in AD. We include 154 participants (78 healthy subjects, 76 patients with AD) from two different languages (106 English speaking and 47 French speaking). Each participant completed a picture description task, in addition to a battery of neuropsychological tests. Each response was recorded and manually transcribed. From this, task-specific, semantic, syntactic and paralinguistic features are extracted using NLP resources. Using inferential statistics, we determined language features, excluding task specific features, that are significant in both languages and therefore represent "generalizable" signs for cognitive language impairment in AD. In a second step, we evaluated all features as well as the generalizable ones for English, French and both languages in a binary discrimination ML scenario (AD vs. healthy) using a variety of classifiers. The generalizable language feature set outperforms the all language feature set in English, French and the multilingual scenarios. Semantic features are the most generalizable while paralinguistic features show no overlap between languages. The multilingual model shows an equal distribution of error in both English and French. By leveraging multilingual statistics combined with a theory-driven approach, we identify AD-related language impairment that generalizes beyond a single corpus or language to model language impairment as a clinically-relevant cognitive symptom. We find a primary impairment in semantics in addition to mild syntactic impairment, possibly confounded by additional impaired cognitive functions.

Keywords: Alzheimer's disease, dementia, spontaneous speech, language impairment, picture description, natural language processing, explainability, multilingual machine learning

INTRODUCTION

Alzheimer's disease (AD) is a pervasive neurodegenerative disease that affect millions worldwide and is the most recognizable through its primarily cognitive syndrome—dementia. From 2008 to 2018, over 200 medical trials failed to develop a cure for AD dementia (Ferreira et al., 2018) emphasizing that early detection and intervention is still the best course for managing AD.

AD dementia is most prominently associated with heterogeneous and broad cognitive impairment; the typical and earliest-observable hallmarks are impaired memory and executive functions (Buckner, 2004). However, language impairments have been reported occurring in preclinical AD as well as mild, moderate, and severe AD dementia (Kempler, 1995; Klimova et al., 2015) possibly providing a window for screening, continuous monitoring and disease management that can help improve quality of life (Taler and Phillips, 2008; Le et al., 2011; Berisha et al., 2015; Klimova et al., 2015). As language is a pervasive aspect of daily living, language-based AD dementia assessment is ecologically valid and, from the patient perspective, one of the least intrusive ways to assess symptoms of AD dementia. This situates language impairment as an interesting behavioral biomarker from both a clinical and patient perspective (Ferris and Farlow, 2013).

Evidence for language impairment in AD dementia stems from studies using a variety of assessments ranging from structured, clinically-validated tasks to unstructured conversation (for an overview, see Szatloczki et al., 2015). An example of a structured task would be a naming task where a person is shown images on cards and asked to name the object. However, naming tasks do not represent the structure or nuance of natural language. In comparison, an unstructured clinical interview between a clinician and patient produces spontaneous speech in its full variance but is difficult and costly to compare and evaluate for minimal changes in cognition, including language, on a qualitative level. Therefore, many reported studies use a standardized experimental setup to elicit spontaneous speech from subjects; often, this is done by picture description tasks (for an overview, see Mueller et al., 2018). In the middle of this spectrum, the picture description task is a clinically-validated task where a patient is asked to describe a standardized picture. This produces spontaneous speech about an anticipated set of topics that is comparable among populations.

With an emphasis on available picture description data, AD detection has been a popular field for applied automatic speech processing and advanced natural language processing (NLP). The goal of such studies is to ultimately discriminate between a form of dementia and healthy control subjects (HC). In a fully automatic system, an audio recording is automatically transcribed with automatic speech recognition (ASR; König et al., 2015). This creates two sources of information from the file: (1) the sound recording; and (2) the text transcription. To model these sources of information, features are either implicitly represented (Orimaye et al., 2014) or explicitly engineered to automate clinical qualitative analysis (Fraser et al., 2016) and extracted from both components of the task. These features are then used to train supervised machine learning (ML) classifiers to discriminated conditions between a pathological patient group and healthy subjects (Yancheva et al., 2015; Yancheva and Rudzicz, 2016; Fraser et al., 2019).

These recent computational approaches represent significant advances for a better understanding of the AD dementia-related language impairment and including the technical challenge to efficiently assess spontaneous speech, but we argue that there are still multiple caveats. With advanced computational techniques and ML methods, there is an increased complexity added to understand the classifiers' decisions and the entailed clinical assumptions. In other words, good ML performance alone does not necessarily entail clinical evidence for language impairment as a cognitive symptom in AD dementia. Additional methodological precautions must be taken to ensure that findings are clinically-valid, generalizable and do not over fit to a single corpus or language. Hence, limitations in current research have been attributed to lacking standardization and comparability between diagnostic settings as well as a growing gulf between how computational features actually model clinically-observable change (de la Fuente Garcia et al., 2020). The result being a lack of translation between NLP research and clinical application.

We state, that a major research gap is present between the clinical understanding of language impairment (as a neurocognitive function impairment) apparent in everyday spontaneous speech and recent NLP techniques used together with ML for speech-based classification of AD. To overcome this, we will: (1) investigate automatically extracted NLP features from spontaneous picture descriptions with respect to their ability in robustly capturing clinically valid AD-related language impairment; and (2) train robust ML models capturing cognitive language impairment in AD with afore-identified generalizable and explainable NLP features.

BACKGROUND

In order to model language impairment in AD, we first investigate which subprocess of language are impaired as defined by clinical literature. Language impairment in AD dementia is characterized by declining semantic and pragmatic processes and reduced syntactic complexity. Semantic processes refer to the meaning of language. A reduction in semantic processes is often indicated by difficulty finding a specific word, loss of comprehension, finding the incorrect word, using ambiguous referents, creating new words, and loss of verbal fluency. Pragmatic processes refer to adapting language to a specific situation. Pragmatic deficits can result in a person with AD dementia language impairment speaking too loudly, speaking at in appropriate times, repeating themselves or digressing from the topic. Syntactic processes are associated with the underlying structure of language and sometimes grouped together with grammaticality. In early stages, syntactic processes and speech processes remain preserved (Savundranayagam et al., 2005; Ferris and Farlow, 2013; Klimova et al., 2015). However, complexity of syntax in written language has been shown to be significantly associated with cognitive impairment (Aronsson et al., 2020). In addition, ML classification experiments have identified syntactic impairment in the AD Dementia groups (Fraser et al., 2016). Beyond identifying known language impairment, it is crucial to consider that speech and language processes do not occur in isolation and are intertwined with other cognitive and physical processes.

Impaired Language vs. Impaired Speech

Impaired speech is the physical process of speaking involving the lungs, trachea, vocal chords and mouth whereas impaired language refers to deficits in the cognitive process of forming language with structure and meaning. While ML approaches are a powerful tool to estimate the utility of spontaneous speech features, interpreting them in a neuropsychological sense remains challenging. Although speech features are extracted from spoken language, this does not necessarily entail that they reflect language as a neurocognitive function as speech is confounded with multiple neurocognitive processes as well as gender, age and culture. As a result, not all well discriminating speech features can be assumed as evidence for the cognitive aspects of language deficits in AD dementia.

Compound Cognitive Processes and the Picture Description Task

Cognitive, language, and speech processes are interdependent employing multiple aspects of cognition: retrieval from semantic and episodic memory, sustaining and dividing attention for error monitoring, as well as working memory for syntax production (Mueller et al., 2018). For instance, inability to recall a specific word—a semantic deficit—can result in a person with AD not being able to maintain concentration on the task—a pragmatic issue (Ferris and Farlow, 2013).

Since spontaneous descriptions of pictures are a compound cognitive performance of multiple neurocognitive functions and do not purely represent language impairment, when modeling impaired language processes embedded in speech, additional theoretical guidance and architecture within the ML experiments are needed to interpret speech-based features. It is not safe to assume that all well-discriminating ML features (in an AD vs. HC setup) are intuitively explainable, or even relevant, with respect to underlying cognitive processes. Spontaneous speech from the picture descriptions task is a compound of cognitive functions including language. Therefore, careful feature curation is needed to ensure that features are truly measuring language impairment and not just task performance.

Natural Language Processing and the Picture Description Task

Most qualitative analyses of spontaneous speech picture descriptions try to model cognitive impairment by leveraging a variety of computationally extracted features. Calz et al. (2021) reviewed 51 studies for dementia detection from the very common Cookie Theft Picture Description Task (CTP; Goodglass et al., 2001), collected and split 87 features into: rhythmic, acoustic, lexical, morpho-syntactic, and syntactic subgroups. Fraser et al. (2016) engineered features and categorized them into: part of speech, syntactic, grammatical constituency, psycholinguistics, vocabulary richness, information content, repetitiveness, and acoustic subgroups. Using factor analysis, they conclude on findings of semantic impairment, syntactic impairment, information impairment, and acoustic abnormality. For our analysis, we build off this finding to create four feature subsets: task-specific, semantic, syntactic and paralinguistic features (see also **Figure 1**). While it is arguably impossible to fully disambiguate each feature into a single category (Savundranayagam et al., 2005; Ferris and Farlow, 2013), we argue to evaluate features based on the following structure.

Task-Specific Features

In clinical practice, the CTP task is scored by counting the number of unique entities that a person mentions in the picture, referred to as information units (IUs). The individual counts of IUs in the CTP task (e.g., the number of times someone says cookie) are often used in automatic classification scenarios for cognitive impairment (Zraick et al., 2011; Fraser et al., 2016, 2019; Eyigoz et al., 2020). However, we argued that these individual counts are not indicative of semantically-motivated language impairment but rather represent task-specific performance or task completion. This is underpinned by the finding that most of the individual IU count features are not correlated with other classic psychometric language function assessments (Kavé and Goral, 2016). Fraser et al. (2016) found that including these features in ML experiment could be explained by information impairment as well as semantic impairment and represents a joint effort of multiple neurocognitive functions. In addition, IU count-based features are currently recognized as being task-specific also in state-of-the-art work on this topic (Robin et al., 2020). Thus, these features are treated as a measurement of general task performance in this study and not as indications of language impairment.

Semantic Features

It is generally accepted that one of the earliest characterizable impairments caused by AD dementia are semantic processes (Appell et al., 1982; Martin and Fedio, 1983; Bucks et al., 2000; Savundranayagam et al., 2005; Ferris and Farlow, 2013; Klimova et al., 2015). When modeling semantics, features are engineered

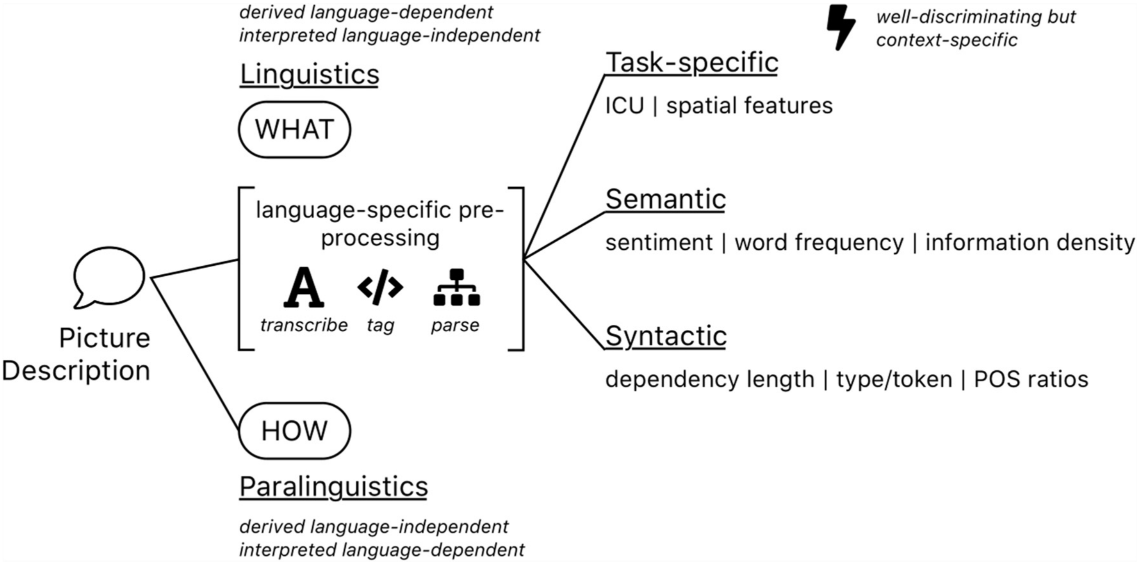

derived language-dependent
interpreted language-independent
Linguistics
WHAT
well-discriminating but
context-specific
Task-specific
ICU | spatial features
language-specific pre-processing
A
transcribe
tag
parse
Semantic
sentiment | word frequency | information density
Picture Description
Syntactic
dependency length | type/token | POS ratios
HOW
Paralinguistics
derived language-independent
interpreted language-dependent

FIGURE 1 | A schematic overview of feature kinds that are typically extracted from spontaneous speech picture descriptions. Some of them involve extensive pre-processing steps such as automatic speech recognition (ASR), part of speech tagging or sentence parsing and additional linguistic resources for calibration, others not.

to capture what is being said. In the CTP task, the semantics are constrained to what is happening in the image, allowing features to be extracted in an automatic and anticipated fashion. Here, semantic features are defined in the CTP as the high-level grouping of named IUs, commonly used by clinicians use to evaluate the task, and not the individual count of each IU. As an example, the number of times the patient says "girl" is not a generalizable representation of semantics but the total number of named IUs in the image can be used to measure ability to explore the semantic space. It has been shown that semantic measures, usually implemented in predefined IUs that represent the content of the to-describe picture, yield across the board good results in classifying between AD dementia and HC (for a review, see Mueller et al., 2018). Previous studies have reported that the AD group reports generic IU features (e.g., girl) without exploring more specific terms (e.g., sister, daughter; Eyigoz et al., 2020). We expect semantic impairment to be prevalent and evident between corpora and languages.

Syntactic Features

In this automatic scenario, syntactic features are engineered to represent the structure of language. This can manifest in a quantifiable way such as differences of sentence complexity or increased use of certain parts of speech. Other studies have reported significant AD dementia-related language impairments from picture descriptions as measured by syntactic features (Lyons et al., 1994; Kempler et al., 1998; Ahmed et al., 2013; Fraser et al., 2016; Yancheva and Rudzicz, 2016). This representation of language requires language specific resources in order to be calculated. We hypothesize these features to be moderately language dependent but some features to represent syntactic impairment that overlaps between languages.

Paralinguistic Features

Paralinguistic features—sometimes also referred to as acoustic, audio or speech features—are specifically appealing for automated speech analysis as they require minimal to no pre-processing and in theory capture the full variance of the acoustic signal and therefore the pathological speech behavior. The calculation of the features is often borrowed and repurposed from ASR systems, where the measures are done on the physical representation of the speech signal. There are multiple examples that successfully use paralinguistic features extracted from spontaneous speech picture descriptions to effectively discriminate between dementia and HC (Pakhomov et al., 2010; Satt et al., 2014; König et al., 2015; Fraser et al., 2016, 2019; Yancheva and Rudzicz, 2016). Due to the limited involvement of error-prone pre-processing steps (e.g., ASR to derive transcripts for further linguistic analysis) the use of paralinguistic features is often regarded as particularly robust and generalizable (Satt et al., 2014). In contrast, other studies found that paralinguistic features are particularly bad at modeling longitudinal trajectory of dementia or predict established clinical staging scores (Yancheva et al., 2015). From a theoretical point of view, we argue that paralinguistic features have great potential to model differences between AD dementia and HC within a certain data set but at the same time bear an equally great risk of over fitting to the particular language or data set. In terms of monitoring language impairment, it is very unlikely a clean proxy for language impairment in AD dementia can be obtained from speech features but at most for other cognitive (attention or executive functions), physical (lung capacity, vocal tract length) or pathological correlates (affective symptoms) associated with AD dementia (Alario et al., 2006; Baese-Berk and Goldrick, 2009; König et al., 2019).

MATERIALS AND METHODS

To investigate explainable and generalizable NLP approaches for automatically classifying between AD related language impairment and healthy controls, implemented the following three-step methodology:

1. First, a multilingual corpus of English and French spontaneous speech picture descriptions is introduced. Then, features are engineered and sorted into subgroups (task-specific, semantic, syntactic, paralinguistic) based on the aforementioned theoretical considerations. For each corpus, an identical set of features are extracted.
2. In a second step, taking advantage of the multilingual corpora, an inspection of cross-language correlations and statistical significance testing is done. Following the idea that well-differentiating features that model generalizable language impairment as a neurocognitive construct should be significant in both languages.
3. To arrive at explainable and generalizable classification results, ML experiments are conducted separately in the two different languages and in a multilingual setting. For each setting, a classification is done among all semantic, syntactic and paralinguistic features. This is compared to classification results where only "generalizable language" features are used. Generalizable language features are defined as semantic, syntactic and paralinguistics features that are significant in both languages.

By leveraging a multilingual approach, we aim to identify AD related language impairment that generalizes beyond a single corpus or language and models the processes of clinically observable language impairment.

Participants

In this article we include 154 participants (78 healthy subjects) from two different languages (106 English speaking and 47 French speaking) drawn from two different available corpora (English, 2020 ADReSS INTERSPEECH challenge and French, EIT-Digital ELEMENT project); for a comprehensive overview of all demographics see **Table 1**.

The English ADReSS sample (Luz et al., 2020) is a balanced (age- and gender-matched) subset of English DementiaBank (Macwhinney et al., 2011) of 53 HC and 54 confirmed AD patients. There are a total of 106 normalized recording and manually annotated transcripts of the cookie theft picture description task. This subset is derived from the DementiaBank corpus, which is part of the larger TalkBank project (Macwhinney et al., 2011). Patients were assessed between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh (for a detailed description of the cohort see Becker et al., 1994). Participants were referred directly from the Benedum Geriatric Center at the University of Pittsburgh Medical Center, and others were recruited through the Allegheny County Medical Society, local neurologists and psychiatrists, and public service messages on local media. Inclusion criteria were as follows: above 44 years of age, at least 7 years of education, no history of nervous system disorders or be taking neuroleptic medication, initial Mini-Mental State Exam (MMSE) score of 10 or greater and had to be able to give informed consent. Participants with dementia had a relative or caregiver acting as an informant. Participants received neuropsychological and physical assessment and were assigned to the "patient" group primarily based on a history of cognitive and functional decline, and the results of a mental status examination. In 1992—after the end of the study—the diagnosis of each patient was confirmed through clinical record and if available autopsy.

The French ELEMENT sample (König et al., 2018) contains 47 participants that completed the cookie theft picture description task. The initial participant pool was 179 subjects but only 47 participants were given the CPT task while the others were given a different spontaneous speech picture description and therefore are not considered in this study. Participants were recruited within the framework of a clinical study carried out for the EIT-Digital project ELEMENT, speech recordings were conducted at the Memory Clinic located at the Institut Claude Pompidou and the University Hospital in Nice, France. The Nice Ethics Committee approved the study. Each participant gave informed consent before the assessment. Speech recordings of participants were collected using an automated recording app which was installed on an iPad. The application was provided by researchers from the University of Toronto, Canada, and the company Winterlight Labs. Each participant underwent the standardized process in French Memory clinics. After an initial medical consultation with a geriatrician, neurologist or psychiatrist, a neuropsychological assessment was performed. Following this, participants were categorized into different groups: control participants (HC) that were diagnosed as cognitively healthy after the clinical consultation and patients that were diagnosed as suffering from Alzheimer's disease and related disorders (AD). For the AD, the diagnosis was determined using the ICD-10 classification of mental and behavioral disorders (World Health Organization, 1992). Participants were excluded if they were not native speakers or had any major hearing or language problems, history of head trauma, loss of consciousness, addiction including alcoholism, psychotic or aberrant motor behavior or were prescribed medication influencing psychomotor skills. Among the 47 participants that performed the CPT, 22 participants were diagnosed with Alzheimer's disease or related dementias (AD) and 25 participants with subjective memory complaints but no detectable dementia. A Kruskal–Wallis H test revealed significant age differences ($\chi^2_{(1)} = 9.79, p < 0.01$) but no significant difference for education level.

Spontaneous Speech Procedure

In both samples (DementiaBank subset and Dem@Care subset) participants completed a comprehensive protocol of assessments of which for this research only the recordings of the Cookie Theft Picture description task are relevant. In both samples, subjects provided informed consent to be recorded while describing the "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983).

In this task, participants are shown a black and white image of a kitchen with multiple on-going antics while being instructed

TABLE 1 | Sample characteristics for English and French samples.

Language	Diagnosis	N (M/F)	Age	Education	MMSE
English *N* = 106	HC	52 (23/29)	66.13 (6.52)	-	29.10 (1.00)
	AD	54 (24/30)	66.76 (6.61)	-	11.06 (5.49)
French *N* = 47	HC	25 (6/19)	75.40 (7.00)	12.80 (2.08)	28.56 (1.42)
	AD	22 (9/13)	81.59 (4.52)	10.91 (3.94)	18.36 (4.29)

Age in years (SD), Education in years (SD) and score on MMSE cognitive screening with a max score of 30 (SD). Abbreviations: HC, Healthy Controls; AD, Alzheimer's disease; MMSE, Mini Mental State Examination.

to "Tell me everything you see going on in this picture." Testing personnel generally is not meant to provide any feedback during the descriptions of the participants. However, in some cases there is interaction recorded if for example the initial response of the patient is unreasonably brief, such as only a single sentence. Recordings had a mean duration of 62.63 s (*SD* = 35.83) sometimes including prompts from the examiner. The English corpus has an average duration of 70.92 s (*SD* = 36.92) and the French corpus has an average duration of 43.95 s (*SD* = 24.82). All recordings are transcribed according to CHAT protocol (Macwhinney, 1991).

Feature Engineering

For each of the four categories defined previously (semantics, syntax, task-specific, and paralinguistic) features were engineered and then calculated using a program written in the Python programming language (Van Rossum and Drake, 2009; Version 3.7). The following section describes the computation of the features by sub-group. If a language-specific resource is used, the equivalent resource is used for each language in the data.

Task-Specific Features (*N* = 107)

Croisile et al. (1996) defined a set of general IUs that appear in the CTP task (e.g., girl, boy) and these IUs are mapped to a larger set of synonymous keywords (e.g., brother, girl). For instance, the boy in the picture may also be referred to as brother or son. This is done for the following IUs: boy, girl, woman, kitchen, exterior, cookie, jar, stool, sink, plate, dishcloth, water, window, cupboard, dish, curtain. A table of the mappings for each IU category to its keywords is provided in the **Supplementary Materials** for both French and English. For each IU, three features are computed: a binary value to see if the IU is mentioned, the count of times the IU is mentioned, and the ratio of the IU to all mentioned IUs. For spatial features, the CTP image is divided into different subgroups[1]. Three divisions of the image are considered: halves, quadrants and vertical stripes. Halves is where IUs are defined as being on the left side or right side. Quadrants breaks the image into four equal squares, north-east, north-west, south-east and south-west. Vertical stripes cut the image vertically into most-left, center-left, center-right and most-right (Goodglass and Kaplan, 1972). For each of the subsections the following features are calculated: word count, type-to-token ratio, keyword-to-word ratio, and percent uttered. For the division in halves, the number of switches between the sides is considered.

Semantic Features (*N* = 20)

Some semantic features utilize task specific resources, but model semantics by combining the defined IUs—and their mapped keywords—into refined, global semantic features rather than counting individual IUs. A table with the mappings between the IU and the keywords that make up the IU are provided in the **Supplementary Materials** for both English and French. Semantic features calculated with the IUs and keyword mappings are defined in **Table 2**. In addition to the features in the table, semantic features that do not rely on the IU definitions are also calculated. The Word Frequency package for python (Speer et al., 2018) is used to determine the mean, median, and max word frequency of all words as well as mentioned keywords. In addition, the mean, median and max word length is calculated for all words as well as the keywords. To gauge lexical richness of the responses, the type-to-token ratio (TTR) is calculated by dividing all unique words said by the total word count. The Moving-Average-Type-Token Ratio (MATTR) is calculated using a fixed window size of 10. For this measurement, a ratio of the number of distinct words in the sliding window is divided by the total count of words. For example, the TTR for words 1–10 is estimated followed by the TTR for words 2–11, then 3–12, and so on. The resulting TTRs are averaged, the estimated TTRs are averaged. Conceptually, the moving-average type–token ratio MATTR (Covington and Mcfall, 2010) calculates the TTR while reducing the influence that the length of the text has on the measure.

Syntactic Features (*N* = 41)

To evaluate syntax, the mean words per sentence, word count and number of sentences are calculated. In addition, Spacy models are used to calculate the mean dependency length, median dependency length, max dependency length (Honnibal and Montani, 2017)[2]. Using Spacy language models, each participant's response is part-of-speech tagged. The count of each tag, as well as the ratio of the POS tag count to total word count are computed. The following tags are considered: Adjective (ADJ), Adposition (ADP), Adverb (ADV), Auxiliary (AUX), Coordination Conjunction (CCONJ), Determiner (DET), Interjection (INTJ), Noun (NOUN), Numeral (NUM), Particle (PART), Pronoun (PRON), Proper Noun (PROPN), Punctuation (PUNCT), Subordinating Conjunction (SCONJ), Symbol (SYM), Verb (VERB, and Other (X). Specific ratios are calculated between nouns (NOUN) and verbs (VERB), pronouns (PRON) and nouns (NOUN), and determiners (DET) and nouns

[1] Implementation based on https://github.com/vmasrani/dementia_classifier

[2] https://universaldependencies.org/u/pos/

TABLE 2 | Explanation of semantic features.

Example: There is a boy. The boy is a brother. He is stealing a cookie. The sister is watching.		
Feature name	**Explanation**	**Example**
Number of Unique IU (num_unique_IU)	The number of unique IU mentioned Higher means they mentioned more IU in the picture	3, boy and cookie, sister
Number of Unique Keywords (num_unique_keywords)	The number of unique keywords mention Higher means they either used more IU and/or used more lexical variety to describe the IU.	*4, boy, brother, sister and cookie*
Number of Total keywords (num_total_keywords)	Counts all mentions of the IU from the mapped keywords. Higher means they said more overall about the image.	*5, boy, boy, brother, cookie, sister*
Unique IU Density (unique_IU_density)	The number of unique IU (num_unique_IU) mentioned divided by the word count	num_unique_IU = 3; Word count = 18 3/18 = 0.1667
Total IU Density (total_IU_density)	Number of total IU (num_total_IU) divided by the word count.	num_total_IU = 5; Word count = 18 5/18 = 0.2778
Keyword to non-keyword ratio (keyword_to_non_keyword_ratio)	$\frac{num\ total\ keywords}{word\ count - num\ total\ keywords}$	num_total_keywords = 5; Word count = 18 5/(18–5) = 0.3846
Unique IU efficiency (unique_IU_efficiency)	The number of unique keywords (num_unique_keywords) divided by the word count.	num_unique_keywords = 4; Word count = 18 4/18 = 0.22
percentage of IU mentioned (percentage_of_keywords_mentioned)	The number of unique IU (num_unique_IU) mentioned divided by the total count of all IU words available in the image.	num_unique_IU = 3, all_IU_words = 16 3/16 = 0.1875

Keyword Type Token Ratio (keyword_TTR)	The number of unique keywords (num_unique_keywords) divided by the number of total IU (num_total_IU) mentioned.	num_unique_keywords = 4;
		num_total_IU = 5
		4/5 = 0.8

total IU efficiency (total_IU_efficiency)	Number of total IU (num_total_IU) divided by the duration in seconds of the participant's response.	num_total_IU = 5; duration = 15 s 5/15 = 0.33
unique IU efficiency (unique_IU_efficiency)	The number of unique IU (num_unique_IU) divided by the duration in seconds of the participant's response.	num_unique_IU = 3; duration = 15 s 3/15 = 0.2

Feature name contains the name of the feature in the text and the name of each feature use in images in parentheses. The explanation column explains how the feature is calculated. At the top of the table there is an example which is used in the example column to explain how each feature is calculated.

(NOUN). The open (ADJ, ADV, INTJ, NOUN, PROPN, VERB) to closed (ADP, AUX, CON, DET, NUM, PART, PRON) class ratio is also computed.

Paralinguistic Features (N = 208)

To extract paralinguistic features from the normalized wav files free, open-source python libraries, and praat (Boersma and Weenink, 2009) are used.

To characterize the temporal and content features of speech, the My Voice Analysis package[3] is used. This package is developed by the Sab-AI lab in Japan to develop acoustic models of linguistics. This package interfaces the speech analysis research tool praat (Boersma and Weenink, 2009) with python, allowing the following features to be extracted from the wav recording: speech rate, syllable count, rate of articulation, speaking duration, total duration, pronunciation *posteriori* probability percentage score, and ratio of speaking to non-speaking. This package is also used to extract some prosodic features, specifically the mean, standard deviation, minimum, maximum, upper and lower quartile of the F0 value, or what is sometimes referred to as the pitch, in Hertz (Hz).

To represent the sound wave itself, features are borrowed from the ASR community using the Python Speech Features library. The original sound recording undergoes a series of transformations that yield a representation of the sound called the Mel Frequency Cepstrum (MFC). The MFC describes two crucial points of information from the voice to human anatomy; the first is the source (e.g., the lungs) and the second is the filter (e.g., place of articulation). The first transformation separates the source and filter from the signal and then maps this to the Mel scale which approximates the sensitivity of the human ear (Fraser et al., 2018). Typically, up to the first 14 coefficients are used as they represent the lower range frequencies of the vocal tract and yield most of the information (Hernández-Domínguez et al., 2018). This has been shown to be effective at identifying AD patients in previous literature (Dessouky et al., 2014; Rudzicz et al., 2014; Satt et al., 2014; Fraser et al., 2018; Panyavaraporn and Paramate, 2018; de la Fuente Garcia et al., 2020; Meghanani and Ramakrishnan, 2021). From this new representation, the first 14 coefficients of the MFC are extracted and the mean, variance, skewness and kurtosis are calculated for the energy (static coefficient), velocity (first differential), and acceleration (second differential). These are also calculated for the velocity and acceleration, where velocity is the difference between consecutive time steps, and acceleration is the difference between consecutive time steps for each velocity. Additionally, the mean, maximum, minimum and standard deviation of the root mean square value (RMS), centroid, bandwidth, flatness,

[3] https://pypi.org/project/my-voice-analysis/

zero crossing rate (ZCR), flatness, loudness, and flux of the spectrogram are calculated with the Librosa[4] package.

Inferential Statistical Analysis

After extracting identical feature sets from both corpora, features are evaluated with regard to their significance in differentiating between the two groups (AD and HC) using non-parametric group comparison and correlation analysis.

Significance Testing

For group comparisons, a non-parametric Kruskal–Wallis H-test for significance is done for each feature to test for significant group differences between the HC and AD samples. Due to the number of performed significance tests, we also report a Bonferroni adjusted probability. This is done separately for each language, meaning each feature has four significance values: English p-value, English adjusted p-value, French p-value, and French adjusted p-value. Significance was set at $p < 0.05$.

Correlation Analysis

Correlation analysis was used to arrive at a continuous numeric variable describing the ability of a feature in discriminating between AD and HC (AD/HC × feature value) which is at the same time comparable between both languages/samples; this is mainly relevant for plotting the discriminative power of feature in both languages and better visualizing the generalizability of the extracted features. For correlation values, a point-biserial correlation is calculated between each feature and the nominal group condition.

Machine Learning Experiments

For all ML experiments, we investigate three classifiers: a classic logistic regression (LR) with an L2 regularization, a Support Vector Machine Classifier (SVM), and a simple neural approach with a multilayer Perceptron (MLP) using a logistic activation function and the regularization term (alpha) set to 0.01. All other parameters are left at their default setting. Due to the small size of the data sets in this article, we opted to maximize the available data using leave one out cross validation. For this method, one sample is held for testing and all other data points are used for training. This is repeated so that every sample in the data has been held out one time. While leave-pair-out cross validation is considered to be a less biased approach for binary classification because it exhaustively tries every possible combination, leave-one-out cross validation is a common training-testing split in this line of research (Cohen and Pakhomov, 2020; de la Fuente Garcia et al., 2020; Luz et al., 2020). Even on very small datasets, leave-pair-out cross validation is computationally expensive (Maleki et al., 2020). In order to keep our work comparable with prior and future studies, we opted to use leave one out cross validation as the best method for maximizing the available data while reducing training bias and maintaining reproducibility (Pahikkala et al., 2008; Fraser et al., 2019; Maleki et al., 2020).

Reported scores are the average across all iterations of the classification experiment. All ML experiments are implemented using the python library, scikit-learn[5] (Pedregosa et al., 2011).

Selecting Generalizable Features

To determine which features capture language impairment that is not corpus-specific, the uncorrected Kruskal–Wallis significance testing described previously in statistical analysis ("Significance Testing" section) is used. Features are selected from each subgroup if they were found to be significant ($p < 0.05$) in both French and English and added to the "generalizable language" feature set. Task-specific features are excluded. The "generalizable language" features are listed in **Table 3**.

Experiment Scenarios

Thus far, we have presented two datasets, French and English ("Participants" section, **Table 1**). By concatenating these two datasets, we generate a third multilingual dataset. In addition, two feature groupings have been proposed; Language features defined as all features in the semantic, syntactic and paralinguistic features [for reference see "Semantic Features (N = 20)," "Syntactic Features (N = 41)" and "Paralinguistic Features (N = 208)" sections, and **Figure 1**] and a subset of these features that are considered to be the generalizable language feature set ("Selecting Generalizable Features" section).

To investigate the performance of the generalizable language feature set, six experimental scenarios are conducted in a binary classification scenario (HC vs. AD). For the first three experiments, English, French and multilingual models are trained using all language features. For the next three experiments, English, French and multilingual models are trained using the generalizable language features. We then compare the performance of the language feature set and the generalizable language feature set to see if the generalizable features help or hurt classification performance.

Establishing a Baseline

To relate these experiments to previous work, we train a baseline model that uses all feature subgroups (semantic, syntactic, task-specific and paralinguistic) in a classification with the previously described English dataset. This situates our methods and results in comparison to the recent ADReSS challenge at Interspeech 2020. The goal of this challenge was to use spontaneous speech picture descriptions to differentiate between AD and HC.

In addition to the experimental scenarios and baseline, we create a baseline classification experiment using only age to consider the affects that the unmatched French population has on the multilingual ML experiment.

Evaluation

For classification performance, Area Under the Receiver Operator Curve (AUC) is reported for each experiment scenario described in "Experiment Scenarios" section. Confusion matrices (Bateman et al., 2012; König et al., 2018) are reported for the multilingual model with the generalizable language feature set. A matrix is reported for the overall classification and then

[4] https://github.com/librosa/librosa

[5] scikit-learn version 0.23.2

TABLE 3 | Statistics as per feature set and language.

	Semantic features											
	English						French					
Feature	r_{PB}	m_{HC}	m_{AD}	χ^2	p	$p_{corr.}$	r_{PB}	m_{HC}	m_{AD}	χ^2	p	$p_{corr.}$
keyword_to_non_keyword_ratio	−0.43	0.16	0.11	19.3	***	***	−0.42	0.15	0.11	8.1	**	0.10
max_word_frequency_IU	−0.27	0.0003	0.0003	7.5	***	0.13	−0.34	0.0003	0.0002	5.3	*	0.47
mean_word_frequency_all	−0.37	0.0089	0.0072	14.7	***	***	−0.38	0.0075	0.0066	6.8	**	0.20
num_unique_IU	−0.60	10.94	6.87	37.8	***	***	−0.64	10.24	5.50	18.6	***	***
num_unique_keywords	−0.60	11.83	7.26	37.3	***	***	−0.60	11.40	5.73	16.6	***	**
percentage_of_keywords_mentioned	−0.60	0.10	0.06	37.3	***	***	−0.60	0.07	0.04	16.6	***	**
total_IU_density	−0.42	0.14	0.10	18.7	***	***	−0.36	0.14	0.11	5.9	**	0.34
total_IU_efficiency	−0.54	0.26	0.16	30.2	***	***	−0.46	0.34	0.21	9.8	**	*
num_total_keywords	−0.43	15.42	10.46	19.3	***	***	−0.57	13.44	6.41	15.1	***	**
unique_IU efficiency	−0.56	0.18	0.10	32.4	***	***	−0.40	0.27	0.17	7.2	**	0.16
unique_IU ratio	−0.45	0.10	0.07	21.2	***	***	−0.34	0.11	0.09	5.5	*	0.43

	Syntactic features											
	English						French					
Feature	r_{PB}	m_{HC}	m_{AD}	χ^2	p	$p_{corr.}$	r_{PB}	m_{HC}	m_{AD}	χ^2	p	$p_{corr.}$
ADP_count	−0.28	7.83	5.39	8.0	**	0.20	−0.50	14.44	6.95	11.4	***	*
ADP_ratio	−0.34	0.06	0.05	12.3	***	*	−0.51	0.13	0.09	11.9	***	*
AUX_ratio	−0.35	0.10	0.09	12.7	***	*	0.30	0.04	0.06	4.1	*	1.00
DET_count	−0.26	17.35	13.52	7.3	**	0.31	−0.45	17.72	9.95	9.4	**	0.09
DET_ratio	−0.43	0.15	0.12	19.3	***	***	−0.32	0.17	0.14	4.7	*	1.00
NOUN_count	−0.34	21.12	15.59	11.9	***	*	−0.49	20.76	11.09	11.0	***	*
NOUN_ratio	−0.48	0.18	0.14	23.8	***	***	−0.38	0.19	0.15	6.7	**	0.42
PRON_ratio	0.25	0.07	0.09	6.4	*	0.51	0.51	0.11	0.18	12.1	***	*
PUNCT_count	0.21	15.15	20.69	4.7	*	1.00	−0.38	1.06	0.32	6.5	*	0.47
PUNCT_ratio	0.36	0.13	0.18	13.8	***	**	−0.34	0.01	0.00	5.4	*	0.91

	Paralinguistic features											
	English						French					
Feature	r_{PB}	m_{HC}	m_{AD}	χ^2	p	$p_{corr.}$	r_{PB}	m_{HC}	m_{AD}	χ^2	p	$p_{corr.}$
bandwidth_mean	0.22	2,022.85	2,153.46	5.0	*	1.00	0.32	2,176.25	2,323.75	4.8	*	1.00
energy_skewness	0.23	0.17	0.32	5.4	*	1.00	0.47	−0.26	0.49	10.4	**	0.26
mfcc1_mean	−0.20	−1.87	−4.43	4.4	*	1.00	−0.36	−5.14	−7.87	6.0	*	1.00
mfcc1_skewness	0.23	0.19	0.48	5.8	*	1.00	0.44	−0.31	0.12	8.8	**	0.63
mfcc10_kurtosis	0.23	0.73	1.00	5.5	*	1.00	0.37	0.40	0.67	6.4	*	1.00
mfcc4_kurtosis	0.22	0.92	1.47	4.9	*	1.00	0.41	0.51	1.02	7.7	**	1.00
normalized_loudness_std	−0.34	0.20	0.18	11.8	***	0.12	−0.56	0.23	0.20	14.4	***	*
ratio_speaking	−0.27	0.46	0.37	7.6	**	1.00	−0.57	0.64	0.48	15.2	***	*
speech_rate	−0.28	1.92	1.41	8.5	**	0.73	−0.47	3.12	2.32	10.2	***	0.28

*Point-biserial correlation coefficient r_{PB}, correlating each feature with the nominal group variable (AD, HC), feature means for HC and AD, χ^2 value of the non-parametric Kruskal–Wallis H-test for group differences between AD and HC, p-value and Bonferroni-corrected p-value. Significances: *** <0.001, ** <0.01, * <0.05. Feature names in Bold indicate that they are significant after Bonferroni-correction in both languages.*

the error is broken down by individual language to investigate if the multilingually trained classifier performs equally in both languages.

RESULTS

Results are reported from the two methodological scenarios: inferential statistical analysis and ML experiments.

Inferential Statistical Analysis

Comparing the overall correlation and significance trends in **Figures 2**, **3**, semantic and task-specific features display similar patterns. In general, these features are negatively correlated in both French and English where AD has lower averages than healthy controls. For syntactic and paralinguistic features, both negative and positive correlations are observed. Paralinguistic features show the most language-specific behaviors, where a mild language preference can also be seen in syntactic features, indicated by points that are far from the dashed line.

Following our above-introduced feature categories, we evaluated statistical significance in differentiating between both groups, AD and HC. Of all features calculated, 30% of task-specific, 28% semantic, 39% syntactic features and 65% of paralinguistics features are not significant in either French or English before significance correction. Before correction, 43% of task-specific, 52% of semantic, 24% of syntactic, and 4% of paralinguistic features of the initially extracted features are significant in both French and English (see also **Table 3**).

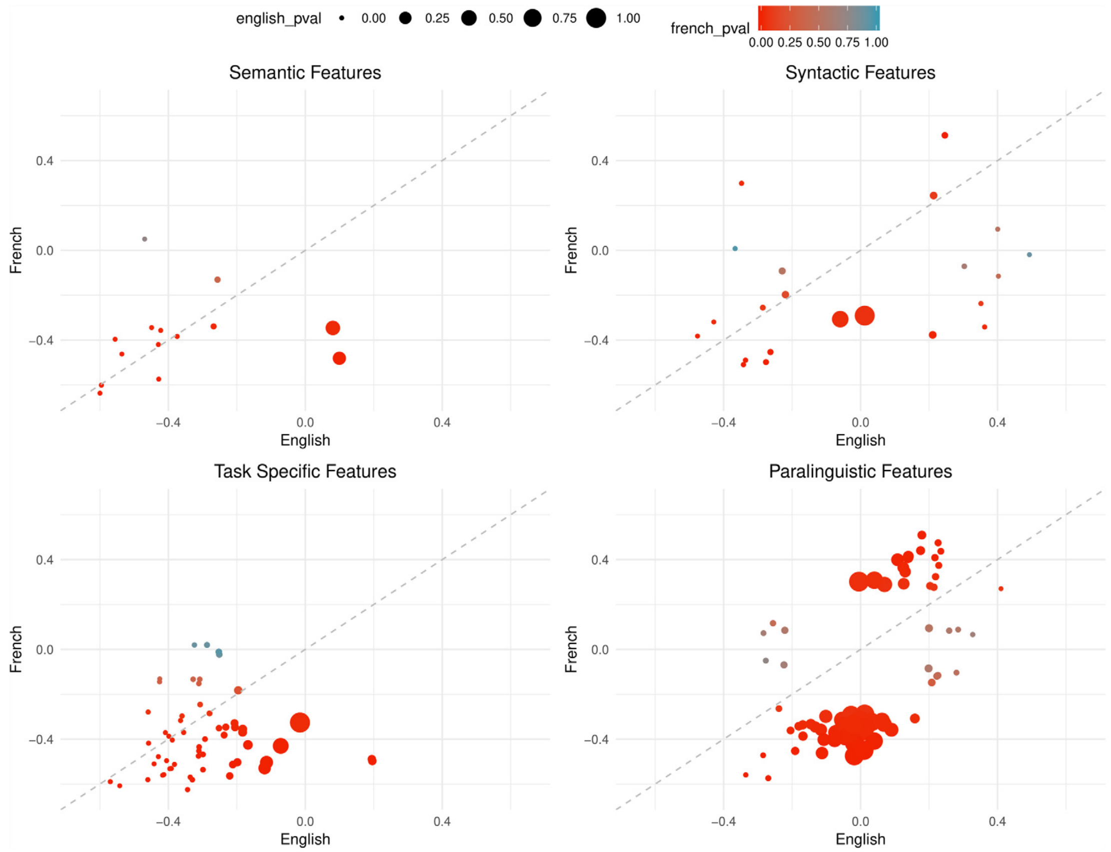


FIGURE 2 | Points are plotted by correlation values (point-biserial correlation coefficient r_{PB}, correlating the feature with the group AD vs. HC) with French on the Y-axis and English on the X-axis for each feature subgroup. The significance value (as by Kruskal–Wallis non-corrected significance test $p < 0.05$) is visualized by point color for French and point size for English. Points closer to the dashed line perform equally well in both languages. This figure contains all features that are significant in EITHER French or English, not necessarily both.

However, due to the large amount features tested ($N_{\text{total}} = 377$), after Bonferroni correction only a fraction of the features remain significant in both languages; 9% task-specific, 24% of semantic, 5% syntactic, and 0% paralinguistic.

Task-Specific Features

Among 107 calculated task-specific features, 32 features are not significant in either French or English, roughly 30%. With significance correction, 75 features are significant in either French or English; 46 features in Both, 20 features in French-only, and nine features in English-only. After significance correction, 10 features remain significant for both languages, approximately 9% of all task-specific features.

Semantic Features

Among 21 calculated features, 15 features are significant in either French or English; two features in French-only, two features in English-only and 11 features in both. While the semantic subgroup has the least calculated features, it has the highest percentage of significant features (approximately 24%) after Bonferroni correction: number of unique IU, number of unique keywords, percentage of keywords mentioned, total IU efficiency, and number of total keywords. For all the significant features in English and French, the AD condition shows lower averages in comparison to the control group (HC).

Syntactic Features

In either language, 25 of the 41 syntactic features are significant in either French or English; 10 features in both, two features in French-only, and 13 features in English-only. After significance correction, noun count and adposition ratio are significant in both languages. For both features, the AD group shows lower averages than healthy controls.

Paralinguistic Features

In either French or English, 72 features among 208 calculated paralinguistic features are significant: nine features in both, 45 features in French-only, 18 features in English-only After significance correction, no features are significant in English and two features are significant in French; ratio of speaking to the full sample duration and the standard deviation of normalized loudness. In both cases, the AD group shows lower averages in comparison to the control averages.

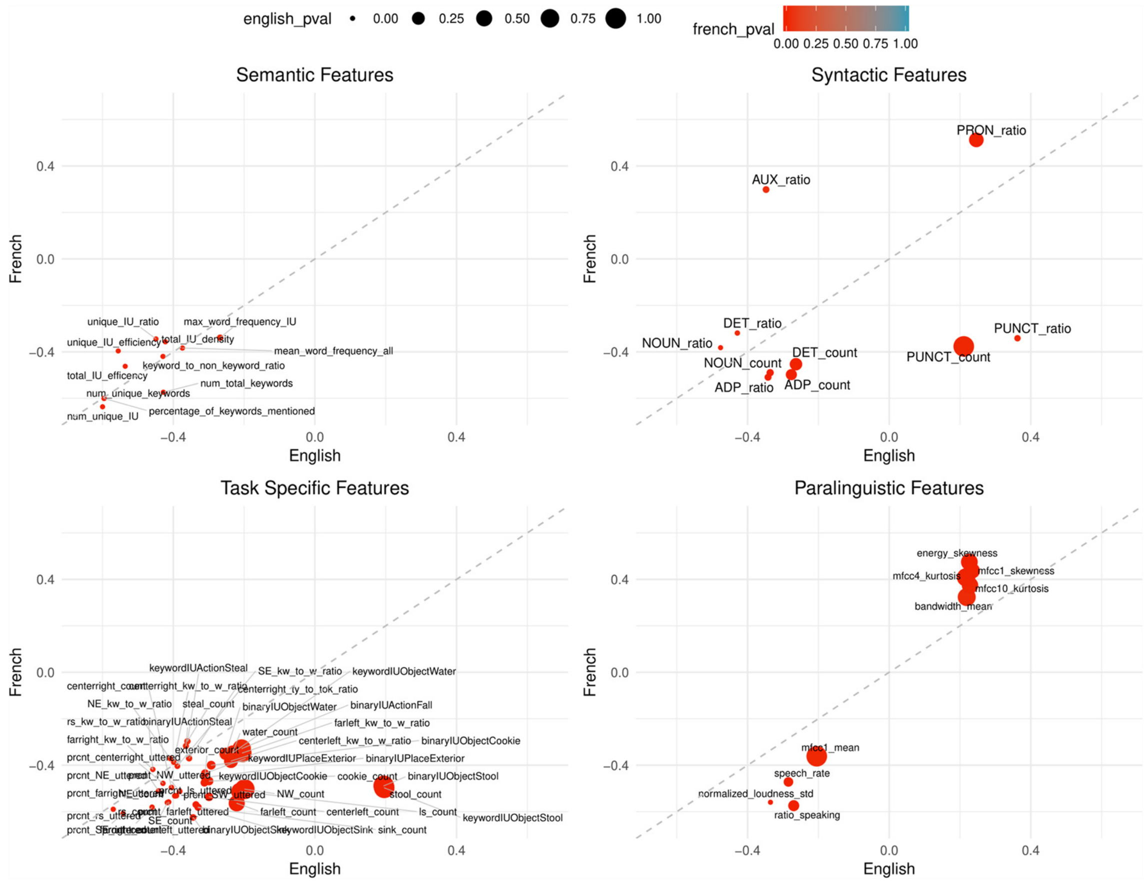


FIGURE 3 | Points are plotted by correlation values (point-biserial correlation coefficient r_{PB}, correlating the feature with the group AD vs. HC) with French on the Y-axis and English on the X-axis for each feature subgroup. The significance value (as by Kruskal–Wallis non-corrected significance test $p < 0.05$) is visualized by point color for French and point size for English. Points closer to the dashed line perform equally well in both languages. This figure contains all features that are significant in BOTH French AND English. Feature labels are added to each point.

In the paralinguistic subplot of **Figure 2**, features are highly polarized as shown by the clustering of points on either side of the dashed line, indicating very little feature overlap between the languages with weaker correlations—especially for English—in comparison to the other feature subsets. By looking at **Figure 3** where features are significant in English and French, lower correlations and the features that are highly polarizes towards one language do not appear in the sub-graph. Among 208 features, only nine features are significant, before correction, and the remaining features are more highly correlated with French than English.

Machine Learning Experiments

Machine learning model performances are visualized for the baseline and experimental scenarios in **Figure 4**. Not included in the graph is the addition classifier for age. All the multilingual experiments were below chance (AUC = 0.5) for age: LR had an AUC of 0.49, SVM had an AUC of 0.38, and the MLP had an AUC of 0.40. This leads us to believe that age is not a good distinguisher between the HC and AD groups for the generalizable experiments. However, it does not eliminate age as a factor from this research and future experiments studies should replicate these findings with age, gender and education balanced data sets to control for possible external conflicting factors.

For the experiments with the LR, all of the models trained with generalizable language features outperform both their respective All language feature models and the English all features baseline. The baseline scenario, the English model with all features, is shown in a solid gray and performs with an AUC of 0.7. English gains 18 points, French gains 23 points, multilingual gains 14 points of AUC over their ALL models. The highest AUC score is nearly tied for the English selected with an AUC of 0.87 and both selected model with an AUC of 0.86. The French select is close in performance with an AUC of 0.85.

For experiments using the SVM, the multilingual generalizable model out-performs all other models reaching an AUC of 0.86. This model improves 10% over the all language feature model. In both English and French, the single language generalizable models outperform their respective all language

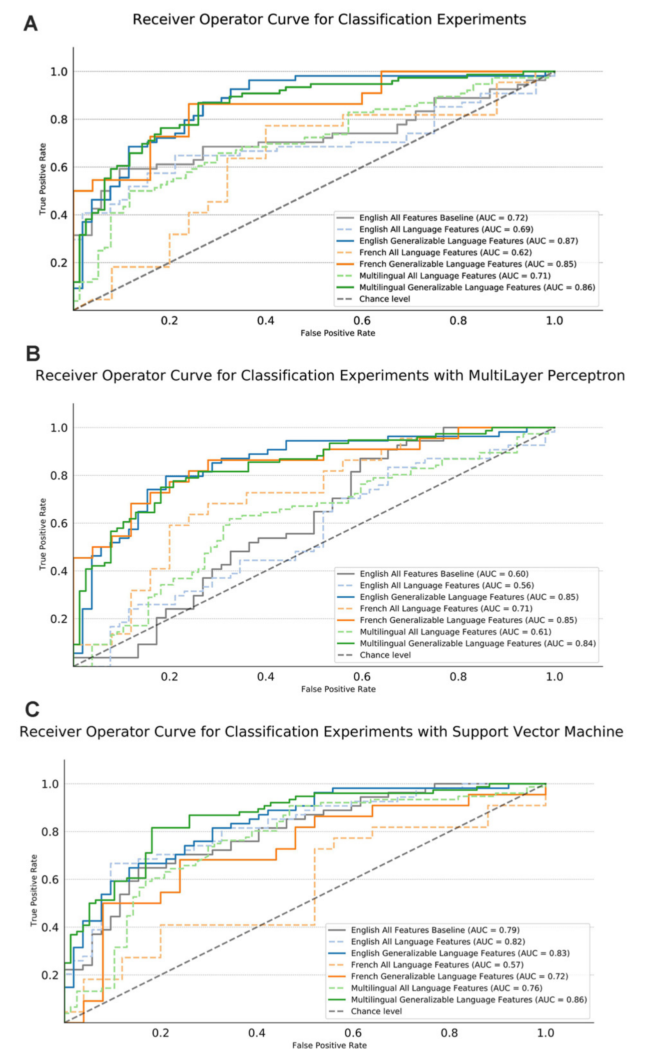


FIGURE 4 | Area Under Curve (AUC) performance results of the machine learning (ML) experiments. English and French for the respective samples separately, multilingual is for the joint classification, multilingual significance testing for feature selection (Generalizable) or using all features (ALL) and using only semantic, syntactic, and paralinguistic features (Language Features). The gray dashed line indicates chance performance of the models. The English (blue), French (orange), and Both (green) models trained with semantic, syntactic, and paralinguistic features are shown with a dashed line. The English, French and Multilingual models trained with the significant, generalizable features in English and French are indicated by the solid lines in the same color, respectively.

feature models. This is more so the case for French where there is an improvement of 15 points, whereas English only improves by 1 point.

For the experiments using the MLP, we see a minor performance drop between the model with ALL baseline features and only generalizable language features. However, in both English and French we see a large performance increase when using only the generalizable features, with both the French and English models reaching an AUC of 0.85. For the multilingual MLP model, we see a mirrored pattern with the LR but slightly lower performance. Overall, we see a 23-point AUC increase when using the generalizable language features in the multilingual scenario, yielding a 0.84 AUC.

We see a similar pattern of error for both the LR, SVM, and MLP models. For the multilingual LR model trained with generalizable language features, the overall error rate is 22.22% and English (22.64%) and French (21.28%) exhibit roughly the same level of error. For the SVM, a slightly lower error of 20.26% is found with a similar split of error between English (20.75%) and French (19.15%). The same result is achieved using the MLP, with a 21.56% overall error and a slightly high, although still comparable, level of the error in both languages (20.75%) English and (23.40%) French.

For both model types, the number of false positives—the case of classifying a control as AD—by language in the multilingual select model make up roughly 30% for French error (27% for the MLP) and 54% for English (59% for the MLP) error. In both models, the English samples have a balanced split of error, but the French model suffers from elevated false positive error. However, this is not the case for the SVM where French (43%) and English (47%) are more balanced in their false positive rate.

DISCUSSION

This article addresses the research gap between the clinical understanding of language impairment (as a neurocognitive functions impairment) apparent in everyday spontaneous speech and recent NLP techniques used together with ML for speech-based classification of AD against healthy control subjects. We propose to: (1) gain insights into AD-related language impairment and its cognitive sub-processes through multilingual NLP feature statistics (generalizing beyond one single language as a cultural phenomenon); and (2) train robust ML models capturing cognitive language impairment in AD with these generalizable features and compare to other methods on the same dataset.

Generalizable NLP Features of Language Impairment in AD

Semantic Features

While the semantic subgroup consists of the lowest number of features, it has the largest number of significant features after Bonferroni correction: number of unique IU, number of unique keywords, percentage of keywords mentioned, total IU efficiency, and number of total keywords. For all the significant features in English and French, the AD condition shows lower averages in comparison to the control group (HC).

Lower averages in number of unique IU, number of unique keywords, number of total keywords, and percentage of keywords mentioned indicates reduced lexical variety and exploration of the available semantic space by the AD group, which is indicative of impaired semantic processes. In addition, there is a reduced semantic efficiency where the AD group is exploring fewer IUs in the same amount of time as controls. For AD patients we found lower overall information efficiency of the uttered descriptions (e.g., total IU density) as well as lower lexical variety with which AD patients described and referred to different IUs in the picture (ratio of unique keywords to all keywords mentioned) in both languages. This is in line with earlier work that finds a decreased semantic efficiency (semantically empty speech) in AD patients' spontaneous speech from picture descriptions (Ahmed et al., 2013; Fraser et al., 2016) but also from other language production tasks (Snowdon et al., 1996; Le et al., 2011).

Overall, semantic features indicate generalizable semantic impairment in AD. Of the feature subgroups, semantic featuresgeneralize the best between the languages, supporting the argument that these features or not task-specific but measure more general semantic abilities. The AD populations, regardless of language, show deficits in semantic scores compare to the health control group.

Syntactic Features

Syntactic features show generally weaker correlations with the pathological state (AD vs. HC) than semantic features. In comparison to the semantic features, syntactic features display a trend of mild language specific behavior (compare also the distance from the dashed line as well as the color/French or size/English of points in **Figure 2**). An interesting finding is the opposing correlation trends for punctuation count and ratio (positive for English and negative for French) and auxiliary ratio (negative for English, positive for French) continuing to indicate that there are syntactic features that are not generalizable for clinical populations because of language. Previous work has shown deficits in determiners, auxiliaries and reduced grammatical structure (Eyigoz et al., 2020). However, the remaining significant syntactic features after correction are adposition ratio and noun count. On average, the AD dementia group use less nouns and adpositions[6].

Adpositions, specifically prepositions, are words used before a noun or pronoun to show time or spatial relationships. For example, in the sentence *the boy is reaching into the cookie jar*, *into* is a preposition showing the spatial relationship of *the boy* to *the cookie jar*. Preposition deficits for AD have been found in Brazilian Portuguese (Alegria et al., 2013). Another study—arguing that spontaneous speech mirroring the decline of effective spatial reasoning in language production—found that AD and HC used the same number of locative/stative prepositions (e.g., in, on, and at) but found significant differences for directional/dynamic prepositions (e.g., into, onto, from, and to; Bosse, 2019).

Although pronoun ratio is only significant in French after the correction, combining this finding with the significant difference in noun count could produce interesting deductions. Between the groups, AD dementia group has a lower average noun count but a greater average pronoun count. Grossman et al. (2007) used new verb acquisition to show that, in comparison to controls, AD dementia patients had fragmented knowledge acquisition. The AD group was able to grammatically use the verb but did not retain its semantic meaning. This could lend insights into the increased pronoun ratio and decrease noun count, where the AD group is not able to recall the semantic names of the IUs in the picture (e.g., boy, brother) and compensates using pronouns (e.g., he). This may be directly related to semantic AD-related language impairment (as described above), where a person uses ambiguous terms (pronouns) instead of specific lexicals (nouns; Savundranayagam et al., 2005; Ferris and Farlow, 2013; Klimova et al., 2015).

While some studies report reduced syntactic complexity in AD patients in earlier detection ML scenarios for the CTP (Fraser et al., 2016), others show contrary findings showing no association between syntactic complexity and cognitive pathology at early stages (Mueller et al., 2018). Evidence from other cognitive tasks show impaired syntax early in disease progression from free spontaneous speech as elicited by questions (Croisile et al., 1996) or written picture descriptions (Kemper et al., 2001).

These findings lead us to believe that syntactic impairment is present but could be confounded as compensation for the profound semantic deficits or other cognitive processes in AD dementia related language impairment.

Paralinguistic Features

For the group of paralinguistic features only around 10% of the initially extracted features were kept after multilingual significance check. Although paralinguistic features are typically reported as important well-classifying features in almost all AD language investigations using computer-aided automatic speech analysis in combination with ML (Pakhomov et al., 2010; Satt et al., 2014; König et al., 2015; Fraser et al., 2016, 2019; Yancheva and Rudzicz, 2016) and explicitly mentioned as robust solutions to the problem (Satt et al., 2014), we find the contrary: the majority of state-of-the-art paralinguistic features do not generalize between languages and therefore are probably not modeling language impairment in AD as a neurocognitive function. Therefore, we argue that they need further clinical investigation to be used as an argument about language impairment in AD.

On the other hand, the question remains why paralinguistic features model differences between healthy and pathological spontaneous speech so well in ML classification scenarios. It could be that they represent variance from other factors such as affective correlates like apathy, which has been shown to affect paralinguistic properties of speech and is a common comorbidity in AD (König et al., 2019), or other non-language neurocognitive functions such as executive functions. For example, we found a lower speech rate in AD patients in both languages which can be interpreted as evidence for a generally impaired psychomotor speed which is highly related with additional factors such as age and executive functions (Keys and White, 2000). However, it is also very likely that from the large amount of extracted paralinguistic features, the "significant" ones just represent statistical artifacts. This can be argued as after Bonferroni correction none of the paralinguistic features yields significance in both languages. This result illustrates well the paradox of paralinguistic features that are highly

[6] A combination of prepositions and postpositions. Postpositions in French seldom occur in spoken language and the only accepted postposition in English is *ago*. Therefore, adpositions, in this application, are assumed to be prepositions.

discriminative in AD vs. healthy control ML experiments but according to traditional interference statistics standards would be considered an artifact suffering from alpha error accumulation. Even without the multilingual generalizability consideration, this methodological paradox typically is disregard in state-of-the-art research combining NLP features for AD classification with ML.

After correction, no features were significant in both languages for English and only two features were significant in French: the ratio of speaking to the full sample duration and the standard deviation of normalized loudness. In both cases, the AD dementia group shows lower average scores than controls. The AD group speaks less overall which can be interpreted as a proxy for overall amount of language production in this task. This possibly reflects semantic, but also multiple other cognitive processes, as previously stated. A lower standard deviation of normalized loudness for the AD group indicates less change in speaking volume as compared to the control group. This could be indicative of common AD-related affective comorbidities such as apathy (König et al., 2019) which result in a less expression and variation in speech patterns.

Overall Findings

Overall, our investigation of generalizable NLP features for language impairment in AD robustly confirms AD patients' semantic impairment in terms of low information efficiency and therefore semantically empty language. This cardinal semantic syndrome can be also additionally confirmed by increased syntactic compensation (using ambiguous terms instead of precise lexical-semantic terminology). Beyond this, we find reduced usage of prepositions, independent of the language, which could be indicative of the earlier-reported decreased complexity in AD language production but more research needs to be done to determine if this is syntactic impairment or confounded by other cognitive processes. Finally, we found almost no paralinguistic features that are indicative of a robust global hence cognitive language impairment in AD except for those who proxy either semantic deficits or affective comorbidities—the latter one indicating a non-causal correlation rather than a robust signal on language impairment in AD.

Machine Learning Models With Generalizable and Explainable Features

Comparison to Baseline

The English baseline classifier with all features (on the same data set as Cummins et al., 2020; Farrús and Codina-Filbà, 2020) achieved an AUC of 0.72 and accuracy of 69.7% using a LR classifier. In comparison, the English classifier with generalizable language features achieved an AUC of 0.87 and an accuracy of 76.4% using a LR model.

On the balanced DementiaBank dataset using both linguistic and paralinguistic features, an 87.5% classification accuracy was achieved using a Random Forest classifier (Farrús and Codina-Filbà, 2020) and an 85.2% using a fusion deep learning approach (Cummins et al., 2020). On a different subset of 167 samples from DementiaBank, combining linguistic and paralinguistic features yielded an 81% accuracy (Fraser et al., 2016).

For multilingual approaches, only semantic word embeddings based on IU features were used to classify in a Swedish and English early detection setting with an 72% accuracy in Swedish and 63% accuracy in English (Fraser et al., 2018). French and English were used to train IU-level language models. The authors report a 0.89 AUC between AD and HC, the best model being trained on both languages (Fraser et al., 2019). The authors could not find any studies where a multilingual approach combined linguistic and paralinguistic features.

Other approaches, not explicitly extracting features, have been used for high performance classifiers on other subsets of the DementiaBank data. Namely, modeling the language of each population and then using perplexity scores (Fraser et al., 2018; Cohen and Pakhomov, 2020) has shown promising results producing interpretable models and reporting AUC scores of 0.93 (Cohen and Pakhomov, 2020). For a more in-depth overview of other methods used for automatic classification used for DementiaBank, please see de la Fuente Garcia et al. (2020).

Model Discussions

Looking at the ML experiments, the multilingual method of feature selection to identify generalizable language features drastically improved every ML performance.

For English, between the baseline with all features and using only language features, there is a small dip in performance when the task-specific features are removed. However, the best English, French and multilingual model performances is with the generalizable language features. More importantly, the performance increase is not only in the multilingual classifier, but a similar level of error is maintained between both languages separately (see **Table 4**). This finding is backed up by the confusion matrices that show a similar distribution of error types across the board. In both languages, as well as in the overall classifier, a comparable number of AD patients were wrongly classified as healthy (false negatives) and a comparable number of healthy subjects got wrongly classified as AD patients (false positives).

It has been shown early on that ML classification of AD and healthy subjects can benefit from a transfer learning approach between multiple languages (Fraser et al., 2019). However, we can show that in spontaneous speech picture descriptions a theory driven and generalizable approach to underlying features not only show good classification results between AD and healthy subjects but at the same time provides clinically-supported evidence of language impairment from spontaneous speech in AD.

Therefore, we conclude that there is evidence of language impairment in AD in everyday spontaneous speech and that this impairment could be driven by a language impairment in the neurocognitive sense. Evidence for this claim is provided by language-independent language impairments as robustly measured by linguistic (semantic and syntactic) and marginally also paralinguistic properties. This is in line with previous research on AD language impairments from traditional clinical research (Kempler, 1995; Taler and Phillips, 2008; Szatloczki et al., 2015).

TABLE 4 | Confusion matrices for the final robust classifier without task-specific features using multilingual significance feature selection.

LR Results			
English and French without task-specific features and feature selection (Error Rate = 22.22%)			
		Ground Truth (Diagnosis)	
		True	**False**
Classification Prediction	AD (positive)	58 (AD/AD)	16 (AD/HC)
	HC (negative)	61 (HC/HC)	18 (HC/AD)
English classifications from the above joint ML scenario (Error Rate = 22.64%)			
		Match to Ground Truth (Diagnosis)	
		True	**False**
Classification Prediction	AD (Positive)	43 (AD/AD)	13 (AD/HC)
	HC (Negative)	39 (HC/HC)	11 (HC/AD)
French classifications from the above joint ML scenario (Error Rate = 21.28%)			
		Match to Ground Truth (Diagnosis)	
		True	False
Classification Prediction	AD (Positive)	15 (AD/AD)	3 (AD/HC)
	HC (Negative)	22 (HC/HC)	7 (HC/AD)
SVM Results			
English and French without task-specific features and feature selection (Error Rate = 20.26%)			
		Ground Truth (Diagnosis)	
		True	False
Classification Prediction	AD (Positive)	59 (AD/AD)	14 (AD/HC)
	HC (Negative)	63 (HC/HC)	17 (HC/AD)
English classifications from the above joint ML scenario (Error Rate = 20.75%)			
		Match to Ground Truth (Diagnosis)	
		True	False
Classification Prediction	AD (Positive)	42 (AD/AD)	10 (AD/HC)
	HC (Negative)	42 (HC/HC)	12 (HC/AD)
French classifications from the above joint ML scenario (Error Rate = 19.15%)			
		Match to Ground Truth (Diagnosis)	
		True	False
Classification Prediction	AD (Positive)	17 (AD/AD)	4 (AD/HC)
	HC (Negative)	21 (HC/HC)	5 (HC/AD)
MLP Results			
English and French without task-specific features and feature selection (Error Rate = 21.56%)			
		Ground Truth (Diagnosis)	
		True	False
Classification Prediction	AD (Positive)	59 (AD/AD)	16 (AD/HC)
	HC (Negative)	61 (HC/HC)	17 (HC/AD)
English classifications from the above joint ML scenario (Error Rate = 20.75%)			
		Match to Ground Truth (Diagnosis)	
		True	False
Classification Prediction	AD (Positive)	41 (AD/AD)	13 (AD/HC)
	HC (Negative)	39 (HC/HC)	13 (HC/AD)
French classifications from the above joint ML scenario (Error Rate = 23.40%)			
		Match to Ground Truth (Diagnosis)	
		True	False
Classification Prediction	AD (Positive)	14 (AD/AD)	3 (AD/HC)
	HC (Negative)	22 (HC/HC)	8 (HC/AD)

The first matrix shows the overall classification result of the model trained on the multilingual data. To ensure this model is not favoring one language, results are further broken down by language in the following matrices. Error is indicated by the false column where a false positive (AD/HC) is the case where a healthy control is classified as having AD and the False negative (HC/AD) is classifying a person with AD as a healthy control. The error rate is reported as all falsely classified participants divided by all participants.

CONCLUSION

This study set out to investigate the robust, generalizable detection of language impairment from spontaneous speech in AD dementia through multilingual ML, with the goal of generating insights between both clinical and NLP researchers.

Based on the proposed methodology, we show possible language impairment in AD in a neurocognitive sense of language that is observable in everyday spontaneous speech. Our approach shows that task-independent language features of AD deteriorated speech point towards neurocognitive language impairments. The primary insights are situated in current clinical understanding of AD dementia related language impairments; There is a theorized primary semantic deterioration but also evidence of a milder syntactic impairment that is confounded by multiple other cognitive processes. In addition, the results support that language impairment could be measured by clinically-motivated NLP techniques without sacrificing overall performance.

The adjacent multilingual feature inspection shows that the feature categories correlate differently between both languages with regard to the significance of their features. This observation is of relevance for the research community interested in detecting language impairment in AD from spontaneous speech picture descriptions because language as a neurocognitive symptom has been found to be impaired in AD for different languages (Ahmed et al., 2013; Szatloczki et al., 2015; Mueller et al., 2018) even though AD itself is heterogeneous in the way it effects individuals (Lam et al., 2013; Ferreira et al., 2018). Hence, we highlight that by catering for explainability and generalizability by design of the ML experiments, research can not only generate efficient clinical applications of NLP methods for AD detection from spontaneous speech but also result in clinically actionable insights.

LIMITATIONS AND FUTURE WORK

The authors would like to acknowledge two main limitations in this study. First, A small clinical data set comes with many challenges. Ideally, to evaluate the ML models, we would use both a training dataset and held-out test set. Unfortunately, this is not available for the French data. Due to the lack of a held-out test set, ML scores could be artificially inflated.

Second, it is possible that poor performance by the paralinguistic features could be confounded by multiple factors: such as gender, the significant difference in age for the French population, and the audio quality of the recordings in DementiaBank. Age and gender have been shown to influence speech patterns and pitch range due to anatomical differences. Future work should investigate what impact these factors has on the explainability and generalizability of paralinguistic features. To support the results in this article, future work should try to replicate this study with more data as well as populations matched by age, gender, and education.

To validate the results presented in this study, future work should investigate this methodology on other clinical tasks that produce spontaneous speech to see if finds hold in more scenarios.

While we used ML to demonstrate that application of generalizable language features, we did not try any optimization techniques to boost results. Future work could look at other classifiers or tuning techniques to improve classification results.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: access to demcare must be granted by the principle investigators; Alexandra König can be reached *via* alexandra.konig@inria.fr. Requests to access these datasets should be directed to Alexandra König, alexandra.konig@inria.fr.

AUTHOR CONTRIBUTIONS

JT and HL contributed equally to this article. AK led the French data collection as well as contributed to the clinical interpretation of the presented approach. All authors contributed to the article and approved the submitted version.

FUNDING

Parts of this research were funded by MEPHESTO project (BMBF Grant Number 01IS20075) and by ATTRACT under Grant Agreement no. 777222 within the Horizon 2020 Framework Programme for Research and Innovation.

ACKNOWLEDGMENTS

We are grateful for technical implementation from Alexandra Mayn and Mario Mina. We would like to the thank Philipp Mueller for proofreading.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2021.642033/full#supplementary-material.

REFERENCES

Ahmed, S., Haigh, A. M., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* 136, 3727–3737. doi: 10.1093/brain/awt269

Alario, F. X., Costa, A., Ferreira, V. S., and Pickering, M. J. (2006). Architectures, representations and processes of language production. *Lang. Cogn. Process.* 21, 777–789. doi: 10.1080/016909600824112

Alegria, R., Gallo, C., Bolso, M., dos Santos, B., Prisco, C. R., Bottino, C., et al. (2013). P4–400: comparative study of the uses of grammatical categories: adjectives, adverbs, pronouns, interjections, conjunctions and prepositions in patients with Alzheimer's disease. *Alzheimers Demen.* 9:P882.doi: 10.1016/j.jalz.2013.08.233

Appell, J., Kertesz, A., and Fisman, M. (1982). A study of language functioning in Alzheimer patients. *Brain Lang.* 17, 73–91. doi: 10.1016/0093-934x(82)90006-2

Aronsson, F. S., Kuhlmann, M., Jelic, V., and Östberg, P. (2020). Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis. *Aphasiology* doi: 10.1080/02687038.2020.1742282

Baese-Berk, M. M., and Goldrick, M. (2009). Mechanisms of interaction in speech production. *Lang. Cogn. Process.* 24, 527–554. doi: 10.1080/01690960802299378

Bateman, R. J., Xiong, C., Benzinger, T. L., Fagan, A. M., Goate, A., Fox, N. C., et al. (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N. Engl. J. Med.* 367, 795–804. doi: 10.1056/NEJMoa1202753

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol.* 51, 585–594. doi: 10.1001/archneur.1994.00540180063015

Berisha, V., Wang, S., LaCross, A., and Liss, J. (2015). Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *J. Alzheimers Dis.* 45, 959–963. doi: 10.3233/JAD-142763

Boersma, P., and Weenink, D. (2009). Praat: doing phonetics by computer.

Bosse, S. (2019). Spontaneous spatial information provided by dementia patients and elderly controls in narratives. *Proc. Linguist. Soc. Am.* 4, 1–9. doi: 10.3765/plsa.v4i1.4463

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology* 14, 71–91. doi: 10.1080/026870300401603

Buckner, R. L. (2004). Memory and executive function in aging and AD: multiple factors that cause decline and reserve factors that compensate. *Neuron* 44, 195–208. doi: 10.1016/j.neuron.2004.09.006

Calz, L., Gagliardi, G., Favretti, R., and Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Comp. Speech Lang.* 65:101113. doi: 10.1016/j.csl.2020.101113.

Cohen, T., and Pakhomov, S. (2020). A tale of two perplexities: sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2005.03593.

Covington, M. A., and Mcfall, J. D. (2010). Cutting the Gordian knot: the moving-average type–token ratio (MATTR). *J. Quant. Linguist.* 17, 94–100. doi: 10.1080/09296171003643098

Croisile, B., Ska, B., Brabant, M. J., Duchene, A., Lepage, Y., Aimard, G., et al. (1996). Comparative study of oral and written picture description in patients with Alzheimer–s disease. *Brain Lang.* 53, 1–19. doi: 10.1006/brln.1996.0033

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., et al. (2020). A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. *Proc. Interspeech* 2020, 2182–2186. doi: 10.21437/Interspeech.2020-2635

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J. Alzheimers Dis.* 78, 1547–1574. doi: 10.3233/JAD-200888

Dessouky, M., Elrashidy, M., Taha, T., and Abd elkader, H. (2014). Effective features extracting approach using mfcc for automated diagnosis of Alzheimer's disease. *Int. J. Data Mining and Knowledge Eng.* 6, 49–59.

Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., and Naylor, M. (2020). Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine* 28:100583. doi: 10.1016/j.eclinm.2020.100583

Farrús, M., and Codina-Filbà, J. (2020). Combining prosodic, voice quality and lexical features to automatically detect Alzheimer's disease. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2011.09272.

Ferreira, D., Wahlund, L. O., and Westman, E. (2018). The heterogeneity within Alzheimer's disease. *Aging* 10, 3058–3060. doi: 10.18632/aging.101638

Ferris, S., and Farlow, M. (2013). Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. *Clin. Interv. Aging* 8, 1007–1014. doi: 10.2147/CIA.S39959

Fraser, K. C., Linz, N., Li, B., Fors, K. L., Rudzicz, F., König, A., et al. (2019). "Multilingual prediction of Alzheimer's disease through domain adaptation and concept-based language modelling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minnesota, MN: Association for Computational Linguistics), 3659–3670. doi: 10.18653/v1/N19-1367

Fraser, K. C., Lundholm Fors, K., and Kokkinakis, D. (2018). Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Comp. Speech Lang.* 53, 121–139. doi: 10.1016/j.csl.2018.07.005.

Fraser, K. C., Meltzer, J., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49, 407–422. doi: 10.3233/JAD-150520

Goodglass, H., and Kaplan, E. (1972). *The Assessment of Aphasia and Related Disorders*. Philadelphia, PA: Lea & Febiger.

Goodglass, H., and Kaplan, E. (1983). *Boston Diagnostic Aphasia Examination Booklet*. Philadelphia, PA: Lea & Febiger.

Goodglass, H., Kaplan, E., and Weintraub, S. (2001). *BDAE: The Boston Diagnostic Aphasia Examination*. Philadelphia, PA: Lippincott Williams & Wilkins.

Grossman, M., Murray, R., Koenig, P., Ash, S., Cross, K., Moore, P., et al. (2007). Verb acquisition and representation in Alzheimer's disease. *Neuropsychologia* 45, 2508–2518. doi: 10.1016/j.neuropsychologia.2007.03.020

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., and Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimers Dement. Diagn. Assess. Monit.* 10, 260–268. doi: 10.1016/j.dadm.2018.02.004

Honnibal, M., and Montani, I. (2017). Spacy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Kavé, G., and Goral, M. (2016). Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. *J. Clin. Exp. Neuropsychol.* 38, 958–966. doi: 10.1080/13803395.2016.1179266

Kemper, S., Thompson, M., and Marquis, J. (2001). Longitudinal change in language production: effects ofaging and dementia on grammatical complexity and propositional content. *Psychol. Aging* 16, 600–614. doi: 10.1037//0882-7974.16.4.600

Kempler, D., Almor, A., Tyler, L. K., Andersen, E. S., and MacDonald, M. C. (1998). Sentence comprehension deficits in Alzheimer's disease: a comparison of off-line vs. on-line sentence processing. *Brain Lang.* 64, 297–316

Kempler, D. (1995). "Language changes in dementia of the Alzheimer type," in *Dementia and Communication*, ed Rosemary Lubinski (San Diego, CA: Singular Publishing Group), 98–114.

Keys, B. A., and White, D. A. (2000). Exploring the relationship between age, executive abilities and psychomotor speed. *J. Int. Neuropsychol. Soc.* 6, 76–82. doi: 10.1017/s1355617700611098

Klimova, B., Maresova, P., Valis, M., Hort, J., and Kuca, K. (2015). Alzheimer's disease and language impairments: social intervention and medical treatment. *Clin. Interv. Aging* 10, 1401–1407. doi: 10.2147/CIA.S89714

König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., and Robert, P. H. (2018). Fully automatic speech-based analysis of the semantic verbal fluency task. *Dement. Geriatr. Cogn. Disord.* 45, 198–209. doi: 10.1159/000487852

König, A., Linz, N., Zeghari, R., Klinge, X., Tröger, J., Alexandersson, J., et al. (2019). Detecting apathy in older adults with cognitive disorders using automatic speech analysis. *J. Alzheimers Dis.* 69, 1183–1193. doi: 10.3233/JAD-181033

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement. (Amst)* 1, 112–124. doi: 10.1016/j.dadm.2014.11.012

Lam, B., Masellis, M., Freedman, M., Stuss, D. T., and Black, S. E. (2013). Clinical, imaging and pathological heterogeneity of the Alzheimer's disease syndrome. *Alzheimers Res. Ther.* 5, 1–14. doi: 10.1186/alzrt155

Le, X., Lancashire, I., Hirst, G., and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary Linguist. Comput.* 26, 435–461. doi: 10.1093/llc/fqr013

Luz, S., Haider, F., De la fuente, S., Fromm, D., and Macwhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2004.06833.

Lyons, K., Kemper, S., Labarge, E., Ferraro, F. R., Balota, D., and Storandt, M. (1994). Oral language and Alzheimer's disease: a reduction in syntactic complexity. *Aging Neuropsychol. Cogn.* 1, 271–281. doi: 10.1080/13825589408256581

Macwhinney, B. (1991). *The Childes Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Macwhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: methods for studying discourse. *Aphasiology* 25, 1286–1307doi: 10.1080/02687038.2011.589893

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., and Forghani, R. (2020). Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clin. N. Am.* 30, 433–445. doi: 10.1016/j.nic.2020.08.004

Martin, A., and Fedio, P. (1983). Word production and comprehension in Alzheimer's disease: the breakdown of semantic knowledge. *Brain Lang.* 19, 124–141. doi: 10.1016/0093-934x(83)90059-7

Meghanani, A., and Ramakrishnan, A. G. (2021). An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech.

Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *J. Clin. Exp. Neuropsychol.* 40, 917–939. doi: 10.1080/13803395.2018.1446513

Orimaye, S. O., Wong, S. M., and Golden (Abuzahra), K. J. (2014). "Learning predictive languistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (ACL 2014)*, eds K. Toutannova and H. Wu (Baltimore, MD: Association for Computational Linguistics (ACL)), 78–87.

Pahikkala, T., Airola, A., Boberg, J., and Salakoski, T. (2008). "Exact and efficient leave-pair-out cross-validation for ranking RLS," in *Proceedings of AKRR*, Espoo, Finland, 1–8.

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. B. (2010). Semantic similarity and relatedness between clinical terms: an experimental study. *AMIA Annu. Symp. Proc.* 2010, 572–576.

Panyavaraporn, J., and Paramate, H. (2018). Classification of Alzheimer's disease in PET scans using MFCC and SVM. *Int. J. Adv. Sci. Eng. Info. Technol.* 8, 1829–1835. doi: 10.18517/ijaseit.8.5.6503.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Robin, J., Kaufmann, L., and Simpson, W. (2020). "Evaluation of speech-based digital biomarkers for Alzheimer's disease," in *Poster Presented at the 13th Clinical Trials on Alzheimer's Disease (CTAD) Conference, digital event*, Boston, MA, USA.

Rudzicz, F., Currie, L., Danks, A., Mehta, T., and Zhao, S. (2014). "Automatically identifying trouble-indicating speech behaviors in Alzheimer's disease," in *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '14)*, (New York, NY, USA: Association for Computing Machinery), 241–242.

Satt, A., Hoory, R., König, A., Aalten, P., and Robert, P. H. (2014). "Speech-based automatic and robust detection of very early dementia," in *Proceedings of Fifteenth Annual Conference of the International Speech Communication Association, Inter-Speech*, Singapore, 2538–2542.

Savundranayagam, M. Y., Hummert, M. L., and Montgomery, R. J. (2005). Investigating the effects of communication problems on caregiver burden. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 60, S48–S55. doi: 10.1093/geronb/60.1.s48

Snowdon, A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the nun study. *JAMA* 275, 528–532.

Speer, R., Chin, J., Lin, A., Jewett, S., and Nathan, L. (2018). LuminosoInsight/wordfreq: v2.2. Zenodo. Available online at: https://doi.org/10.5281/zenodo.1443582. Accessed October 3, 2018.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Front. Aging Neurosci.* 7:195. doi: 10.3389/fnagi.2015.00195

Taler, V., and Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *J. Clin. Exp. Neuropsychol.* 30, 501–556. doi: 10.1080/13803390701550128

Van Rossum, G., and Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: Createspace.

World Health Organization (1992). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines.* Geneva: World Health Organization.

Yancheva, M., Fraser, K. C., and Rudzicz, F. (2015). "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias," in *SLPAT@Interspeech*, Dresden, Germany.

Yancheva, M., and Rudzicz, F. (2016). "Vector-space topic models for detecting Alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2337–2346.

Zraick, R. I., Carr, P. B., Gregg, B. A., Smith-Olinde, L., Ghormley, C., and Hutton, T. J. (2011). Information units produced by persons with mild Alzheimer's disease during a picture description task. *J. Med. Speech Lang. Pathol.* 19, 37–45.

Conflict of Interest: JT and AK own stock in the digtial speech biomarker company ki elements UG (haftungsbeschränkt).

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Computer Science

ORIGINAL RESEARCH
published: 31 May 2021
doi: 10.3389/fcomp.2021.634360

A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning

Natasha Clarke, Thomas R. Barrick and Peter Garrard*

Neurosciences Research Centre, Molecular and Clinical Sciences Research Institute, St George's University of London, London, United Kingdom

Alzheimer's disease (AD) has a long pre-clinical period, and so there is a crucial need for early detection, including of Mild Cognitive Impairment (MCI). Computational analysis of connected speech using Natural Language Processing and machine learning has been found to indicate disease and could be utilized as a rapid, scalable test for early diagnosis. However, there has been a focus on the Cookie Theft picture description task, which has been criticized. Fifty participants were recruited – 25 healthy controls (HC), 25 mild AD or MCI (AD+MCI) – and these completed five connected speech tasks: picture description, a conversational map reading task, recall of an overlearned narrative, procedural recall and narration of a wordless picture book. A high-dimensional set of linguistic features were automatically extracted from each transcript and used to train Support Vector Machines to classify groups. Performance varied, with accuracy for HC vs. AD+MCI classification ranging from 62% using picture book narration to 78% using overlearned narrative features. This study shows that, importantly, the conditions of the speech task have an impact on the discourse produced, which influences accuracy in detection of AD beyond the length of the sample. Further, we report the features important for classification using different tasks, showing that a focus on the Cookie Theft picture description task may narrow the understanding of how early AD pathology impacts speech.

Keywords: machine learning, natural language processing, dementia, connected speech, alzheimer's disease, mild cognitive impairment, discourse, spontaneous speech

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Mohammad Soleymani,
University of Southern California,
United States
Loredana Sundberg Cerrato,
Nuance Communications,
United States

****Correspondence:***
Natasha Clarke
p1607544@sgul.ac.uk

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *27 November 2020*
Accepted: *04 May 2021*
Published: *31 May 2021*

Citation:
Clarke N, Barrick TR and Garrard P (2021) A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning. Front. Comput. Sci. 3:634360. doi: 10.3389/fcomp.2021.634360

INTRODUCTION

Alzheimer's disease (AD) includes a long "pre-clinical" period, during which pathological change accumulates in a patient's brain with no apparent effect on their behavior or performance (Jack et al., 2010). Memory decline often emerges during a period of subtle cognitive alteration known as Mild Cognitive Impairment (Albert et al., 2011), however, disease modifying compounds tested in this prodromal stage have failed to show a treatment effect. Thus, there is a need to identify signs of pathology even earlier (Cummings et al., 2016).

There are two broad approaches to detecting pathology: brief cognitive screening tests and biological markers (biomarkers) of disease. Of the former, the Mini Mental State Examination (MMSE) and Montreal Cognitive Assessment can be administered rapidly (Folstein et al., 1975; Nasreddine et al., 2005), and have reasonably good diagnostic accuracy (AUCs of 85% and 74% for distinguishing MCI from controls (Carnero-Pardo, 2014; Ciesielska et al., 2016)). The detailed Addenbrooke's Cognitive Examination (Hsieh et al., 2013) is more accurate (91% AUC (Matias-Guiu et al., 2016)), but takes longer to administer.

Biomarkers include Magnetic Resonance Imaging (MRI), cerebrospinal fluid (CSF) analysis and Positron Emission Tomography (PET). All three approaches can distinguish AD from controls with accuracies of over 90% (Bloudek et al., 2011), but are less accurate for MCI (Mitchell, 2009; Lombardi et al., 2020). Moreover, they are costly to the healthcare provider and inconvenient for patients, limiting widespread use (Lovestone, 2014; Laske et al., 2015). At the time of writing Amyloid PET is restricted to research use (McKhann et al., 2011).

There is evidence that connected spoken or written language (discourse) begins to change early in the course of AD, possibly prior to MCI (Forbes-McKay and Venneri, 2005; Garrard et al., 2005; Ahmed et al., 2013). Improvements in automated Natural Language Processing have led to the suggestion that computational analysis of connected speech could act as a rapid, low-cost, scalable, and non-invasive assay for early stages of AD (Clarke et al., 2020; de la Fuente Garcia et al., 2020).

A common approach to obtaining a sample of discourse involves the patient describing a scene, such as that depicted in the "Cookie Theft" picture (Goodglass et al., 1983). Standard machine learning algorithms using features automatically extracted from transcripts of the resulting descriptions can classify patients with AD vs. controls with 81% accuracy, while a deep learning approach has achieved similar accuracy in classifying MCI vs. controls (Fraser et al., 2016; Orimaye et al., 2018). Alternative methods of sampling discourse, such as recording unstructured or semi-structured spontaneous speech, have been found to be similarly distinguishable (Garrard, 2009; Berisha et al., 2015; Asgari et al., 2017; Mirheidari et al., 2019).

Another approach involves narration of a learned story (either well-known, such as Cinderella, or a novel narrative presented in pictures)-a cognitively complex task that entails the integration of a story's characters and events within a temporal framework (Ash et al., 2007; Drummond et al., 2015; Toledo et al., 2017). Less well studied is the task of describing a process (such as how to change a tyre). For a review of relevant studies see Petti et al. (2020) and de la Fuente Garcia et al. (2020).

For reasons related to its simplicity, standardization and task constraints, and existence of large volumes of data (particularly the DementiaBank (MacWhinney, 2019)), picture description appears to have largely captured the field of discourse analysis (de la Fuente Garcia et al., 2020). There are, however, significant drawbacks to relying on picture descriptions, including limited richness and length (Ash et al., 2006), the somewhat unnatural nature of the task, and (in the case of the Cookie Theft picture) an outdated depiction of domestic life (Berube et al., 2019). Similarly, procedural recall places constraints on discourse but rarely occurs in everyday conversation and so can result in overly simplified speech (Sherratt and Bryan, 2019). By contrast, conversational speech is instinctive and naturalistic, though without constraints, samples can vary widely in length and content (Boschi et al., 2017). Recall of both overlearned and novel narratives have the potential to produce acceptably long and complex discourse samples, but recollection and engagement may vary.

There have been few formal comparisons of the sensitivities of different speech sampling approaches to early AD. Sajjadi et al. (2012) reported that conversation elicited using semi-structured interviews contained more fillers, (e.g. "er" and "um"), abandoned units (elements of speech that are started but not completed) and grammatical function words than picture descriptions. Conversely, picture descriptions gave rise to more semantic errors, such as substituting the word "dog" for "cat" (Sajjadi et al., 2012). Beltrami et al. (2016) found that a logistic regression classifier showed marginally superior accuracy when trained using acoustic, rhythmic, lexical and syntactic features derived from descriptive discourse compared to two personal narrative tasks (recalling a dream and describing a working day) in an Italian-speaking population of patients with MCI (F1 = 0.78 vs. 0.70 and 0.76). It seems likely, therefore, that the task used to elicit spoken discourse affects not only the accuracy of machine learning classification but also the nature of the features that distinguish patients' discourse from that of controls. Here, we report the accuracy of a series of classifiers using input features automatically extracted from five different speech tasks. We report the features found to be important for classification using the two tasks with the highest accuracy—overlearned narrative recall, and picture description.

MATERIALS AND METHODS

Participants

Fifty participants (see **Table 1**) were recruited from the St George's University Hospitals NHS Cognitive Disorders Clinic: 25 healthy controls (HC) and 25 patients with mild AD (n = 13) or MCI (n = 12) (Petersen criteria (Petersen, 2004)). Diagnoses had been made within two years prior to recruitment using imaging, neuropsychological and/or CSF biomarkers as well as clinical information. HC were either friends and family of patients attending clinic or recruited through the Join Dementia Research system (www.joindementiaresearch.nihr.ac.uk). None of the participants gave a history of other conditions which may affect cognition or language such as stroke, epilepsy or chronic mental health conditions, and all provided informed consent. All spoke English as first language. Ethical approval was granted by the Research Ethics Service Committee London–Dulwich, on November 25, 2016 (ref 16/LO/1990).

Procedures

Global cognition was assessed with the Addenbrooke's Cognitive Examination Third Edition (ACE-III) (Hsieh et al., 2013), a widely used measure scored from 0 – 100, with lower scores representing worse functioning.

Connected Speech Tasks

All tasks were administered by the same individual (NC). Only words spoken by the participant were analyzed. We refer to these different approaches as: Picture Description (PD); Conversational Speech (CS); Overlearned Narrative Recall (ONR); Procedural Recall (PR); and Novel Narrative Retelling (NNR).

TABLE 1 | Participant demographics.

	HC median (IQR)	AD+MCI Median (IQR)	Test	*p* value
Age (yrs)	63 (12)	71 (13)	Mann Whitney *U*	0.018*
Sex (% f)	72%	24%	Chi square	0.001**
Education (yrs)	16 (3.8)	12 (4)	Mann Whitney *U*	0.007*
MMSE (30)	29 (0.70)	24 (2.99)	Mann Whitney *U*	<0.001**

*IQR = interquartile range, MMSE = mini mental state examination, converted from total ACE-III score (Matías-Guiu et al., 2018).** = $p < 0.05$,** = $p < 0.001$.

Picture Description

PDs were elicited using a novel version of the Cookie Theft stimulus, consisting of an updated and colored adaptation the original (Berube et al., 2019). Participants were given the instruction "Tell me everything you see going on in this picture." No time-constraints were imposed.

Conversational Speech

CS was generated using the Map Task (Thompson et al., 1993), in which the participant and the researcher have an A4 map with landmarks depicted, (e.g. "fast flowing river"). The participant's map depicts a route traversing the landmarks with a start and finish point. Acting as "Instruction Giver," they must describe the route to the "Instruction Follower" (the researcher), who recreates the route as faithfully as possible by drawing onto their copy of the map.

Overlearned Narrative Recall

Participants were asked to recall the story of Cinderella from memory. They were given the instruction "I'd like you to tell me, with as much detail as you can, the story of Cinderella."

Procedural Recall

Participants were asked to recount the procedure for making a cup of tea. They were given the instruction "I'd like you to tell me, in as much detail as you can, how you would make a cup of tea."

Novel Narrative Retelling

The wordless picture book "Frog, Where Are You?" (by Mercer Mayer) was used as a stimulus for the generation of a novel narrative. Participants looked through the book once, before describing the story based on the pictures.

Transcription

The resulting sample from each connected speech task was transcribed according to conventions detailed in Garrard et al. (2011). Transcription was completed by a single researcher with a subset of 10% re-transcribed by an independent researcher who was blind to participant diagnosis. The inter-rater reliability for transcription of this sample was 84% based on the Levenshtein distance (Navarro, 2001).

Data Analysis

Linguistic Feature Extraction

Two hundred and eighty-six linguistic features, consisting of fine-grained indices reflecting a range of linguistic and para-linguistic phenomena, were extracted from each connected speech task transcript (**Table 2**). See **Supplementary Material** for full descriptions of features and extraction methods.

Feature Selection

Sparse features (defined as those with > 50% zero values for either class) were removed. To render feature scales invariant values were transformed to a scale between 0 and 1 using the MinMax method. To minimize the danger of overfitting, feature selection was applied in each training fold, using i) feature ranking on mutual information with the class, selecting the top 5, 10, 20, and 40; or ii) logistic regression combined with recursive feature elimination (RFE; Guyon et al., 2002), in which each feature is recursively removed from the set and the regression re-trained to classify groups until the optimal subset of 10 features is found[1].

Machine Learning

Four participant groups were considered: i) those with clinical evidence to suggest the presence of AD pathology, i.e., mild AD plus those with MCI (AD+MCI); ii) MCI alone, iii) AD alone, and iv) healthy controls (HC). Each vector of selected features was used to train a series of linear support vector machines (SVM) to output three binary classifications: HC vs. AD+MCI, HC vs. AD; and HC vs. MCI. SVM have previously been used to achieve good results with similar data (de la Fuente Garcia et al., 2020), and a linear kernel was chosen to enable extraction of coefficients. The value of *C* was set to 100 (as in Fraser et al., 2019).

We calculated accuracy and balanced accuracy, due to class imbalance for subgroup classifications. The latter (**Equation 1**) is similar to conventional accuracy when the classifier performs equally well on either class (or when classes are balanced) but is lower if conventional accuracy is high only due to superior performance on the majority class (Brodersen et al., 2010). TP = true positives, TN = true negatives, FN = false negatives, FP = false positives.

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right). \tag{1}$$

We also report sensitivity [$TP/(TP + FN)$], specificity [$TN/(TN + FP)$], and the AUC. *K*-fold cross-validation was carried out using an 80:20 training:test split; the value of $k = 5$ was chosen to ensure a reasonable sized test set, given the small dataset. Feature scaling and selection was calculated on each training fold and applied to the test fold. The reported performance is an average across the five folds, with standard deviation reported to indicate variability.

Extraction of Important Features

To identify features important for group classification the learnt coefficients, corresponding to weights associated with each

[1] The feature set size of 10 was pre-determined according to the highest average accuracy when using the filter approach: taking the mean accuracy across all five tasks for each threshold of *k*, 10 was the highest.

TABLE 2 | Linguistic domains covered by features extracted from each task transcript (number of features in brackets).

Type	Linguistic feature	Example features
Lexico-syntactic (275)	Word production and complexity (11)	e.g., Mean syllables per word, repeated words
	Parts-of-speech (POS) (18)	% Of POS (e.g., nouns, verbs, coordinates) and ratios (e.g., noun:verb ratio)
	Lexical richness (8)	e.g., Type-token-ratio (TTR; types:tokens), moving average TTR with a window size of 10, 20, 30, 40, and/or 50 if the sample was of sufficient length
	Psycholinguistics (34)	Average normative ratings for e.g., familiarity, concreteness, age-of-acquisition of words
	Psychological processes (50)	% Of words relating to individual psychological processes e.g., anger, time, work
	Syntactic structures and complexity (32)	e.g., mean length of sentence, verb phrases per T-unit (VP/T), complex nominals per clause (CN/C)
	Syntactic parse tree features (4)	e.g., maximum depth, mean depth
	Grammatical constituents (111)	Grammatical constituents of syntax tree e.g., NP—> DT NN, a noun phrase composed of a determiner and a noun
	Shannon entropy (1)	Entropy for letters in the sample (Shannon, 1951)
	Fluency (3)	e.g., false start ratio, filler ratio
	Non-verbal (3)	e.g., pauses, laughter
Semantic (11)	Semantic content (3)	e.g., idea density
	Semantic coherence (9)	e.g., Mean cosine similarity between adjacent sentences utilizing google news word2vec model (Mikolov et al., 2013)

TABLE 3 | HC vs. AD+MCI mean (s.d) SVM classification performance across five-fold cross validation for five connected speech tasks, ranked by accuracy.

Discourse-generating task	Accuracy	AUC	Sensitivity	Specificity
ONR	0.78 (0.08)	0.84 (0.05)	0.75 (0.23)	0.82 (0.21)
PD	0.76 (0.18)	0.84 (0.11)	0.69 (0.30)	0.81 (0.12)
PR	0.74 (0.15)	0.85 (0.19)	0.78 (0.15)	0.74 (0.25)
CS	0.66 (0.11)	0.74 (0.10)	0.62 (0.10)	0.78 (0.31)
NNR	0.62 (0.16)	0.62 (0.10)	0.53 (0.21)	0.72 (0.11)

feature during training, were extracted from each fold and ranked by absolute value. Features selected in only one fold were excluded from further analysis. This method uses information from both the feature selection step and the final training step as an indication of importance and aims to find features that are most stable across the model, thus potentially more generalizable.

Between group analyses were conducted for important features using the non-parametric Mann Whitney *U* test, due to non-Gaussian distribution of features in at least one group. Results were Bonferroni adjusted for multiple comparisons and all *p*-values are reported in their corrected form (significance threshold (α) = 0.05).

Demographic Variables

HC and AD+MCI groups were not balanced for age, sex and years in education (**Table 1**). To explore potential confounding on classification results, important linguistic features from the highest accuracy HC vs. AD+MCI classification were used as input in a linear regression to predict age and education, and a linear SVM to classify sex.

RESULTS

Accuracy of Speech Tasks for Classifying Healthy Controls vs. Alzheimer's disease + Mild Cognitive Impairment

Table 3 shows the classification performance achieved on discourse samples derived from each of the five tasks. ONR, PD and PR produced similar average accuracies and AUCs, but overall sensitivities and specificities varied, with the highest accuracy (0.78) and specificity (0.82) associated with samples generated under the ONR condition.

PD achieved the second highest accuracy (0.76), with similar specificity (0.81) and the same AUC (0.84) as ONR but a lower sensitivity (0.69 compared to 0.75). The condition with the third highest accuracy (PR) achieved the highest sensitivity of all tasks (0.78) but second lowest specificity (0.74). The lowest accuracies and AUCs were obtained using CS and NNR. The s.d. of the mean accuracy and AUC for ONR is smaller than for the remaining tasks (0.08 and 0.05, compared to 0.18 and 0.11 for the second most accurate task, PD) indicating less variability given different training and test data.

Important Features for Classification of Healthy Controls vs. Alzheimer's disease + Mild Cognitive Impairment

In the interests of brevity, we focused on the features important for the two most accurate tasks – ONR and PD – which both utilized multivariate feature selection.

Overlearned Narrative Recall Features

Table 4 shows 12 features, ranked by number of folds and mean rank across all folds, that were selected in at least two cross-

TABLE 4 | Important features of overlearned narrative recall for classifying HC vs. AD+MCI. Ordered by number of folds and then mean rank. Mann Whitney *U* tests Bonferroni adjusted for multiple comparisons and reported in corrected form.

Feature	Linguistic domain	No. folds	Mean rank	Between group comparison			Description
				HC median (IQR)	AD+MCI median (IQR)	*p* value	
BNC spoken freq CW	Psycholinguistics	5	6.4	1.32 (0.28)	1.70 (0.51)	0.001**	Mean frequency rating for content words based on British National Corpus. Higher values = higher frequency
NP –> DT	Grammatical constituents	5	3.2	0.00 (0.00)	0.01 (0.01)	0.294	Noun phrase with a bare determiner e.g., "this," "those"
Entropy	Shannon entropy	5	2.4	4.11 (0.04)	4.07 (0.06)	0.037*	Entropy calculated for letters (Shannon, 1951). Higher values = more information, and less certainty in sequence predictions
PP type rate	Grammatical constituents	4	6.8	0.08 (0.01)	0.05 (0.02)	<0.001**	Rate of prepositional phrases
False starts ratio	Fluency	3	8.7	0.00 (0.00)	0.01 (0.01)	4.605	Ratio of incomplete words
S –> CC NP VP	Grammatical constituents	2	7.5	0.000 (0.00)	0.002 (0.01)	1.173	Sentence with a coordinating conjunction, noun phrase and a verb phrase e.g., "But Cinderella smiled."
Idea density	Semantic content	2	7	0.57 (0.02)	0.54 (0.06)	0.064	Mean propositional idea density per word
Ingest	Psychological processes	2	6	0.13 (0.37)	0.00 (0.00)	0.053	% words that correspond to concept of "ingestion" e.g., hungry, dish
DESWLsy	Word production and complexity	2	5	1.32 (0.04)	1.26 (0.11)	0.043*	Mean number of syllables per word
Health	Psychological processes	2	3.5	0.7 (0.68)	0.00 (0.54)	0.031*	% words that correspond to concept of "health" e.g., clinic, flu
Sixltr	Word production and complexity	2	3.5	14.34 (2.09)	11.76 (5.88)	0.012*	% words longer than six letters
Mean WMD	Semantic coherence	2	2.5	0.88 (0.17)	1.17 (0.49)	0.001**	Mean word movers distance (Kusner et al., 2015) between adjacent sentences, using word2vec (Mikolov et al., 2013). Lower values = greater semantic similarity, and therefore coherence

** = p < 0.05. Features in bold appear important for classification using both picture description and overlearned narrative recall (see **Table 5**).*

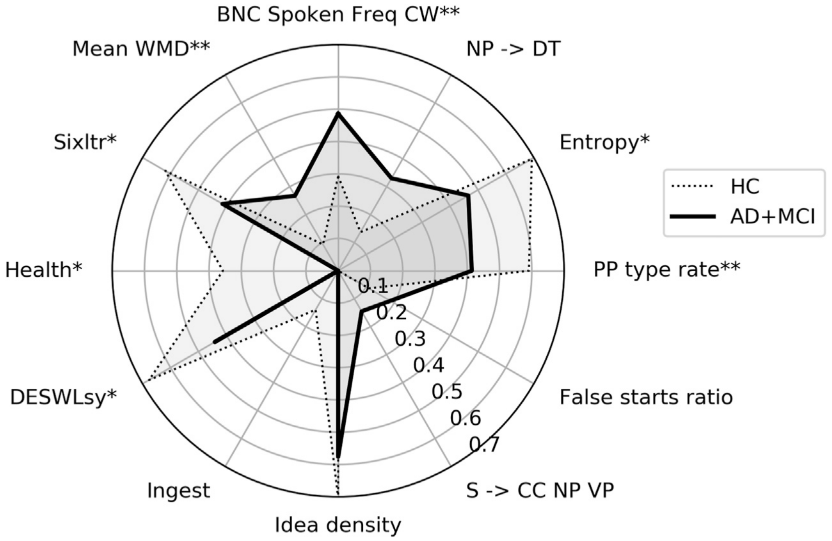


FIGURE 1 | Radar plot showing features important for HC vs. AD+MCI classification using overlearned narrative recall. HC = healthy control, AD+MCI = Alzheimer's disease and Mild Cognitive Impairment group. Features have been scaled to between 0 and 1 using MinMax scaling and medians plotted. * = *p* < 0.05, ** = *p* < 0.001.

TABLE 5 | Important features of picture description for classifying of HC vs. AD + MCI. Mann Whitney *U* tests Bonferroni adjusted for multiple comparisons and reported in corrected form.

Feature	Linguistic domain	No. folds	Mean rank	Between group comparison			Description
				HC median (IQR)	AD + MCI median (IQR)	*p* value	
NP –> DT	Grammatical constituents	5	8.6	0.00 (0.01)	0.01 (0.01)	0.008*	See **Table 4**
Tone	Psychological processes	5	8.4	50.32 (30.84)	32.45 (23.87)	0.021*	Measures overall emotional tone of sample. Higher values = more positive
S – > ADVP NP VP	Grammatical constituents	5	7.2	0.002 (0.01)	0.000 (0.00)	0.042*	Sentence with an adverb phrase, noun phrase and verb phrase e.g., "Hardly anyone noticed."
SUBTLEXus Range FW	Psycholinguistics	4	6.5	8,189.19 (163.97)	8,273.81 (124.38)	0.32	Measures frequency of function words according to their range, (i.e. across documents as opposed to within) using the SUBTL corpus of television and film subtitles
Demonstratives	Parts-of-speech	4	5	0.01 (0.00)	0.01 (0.01)	1.127	Use of demonstratives (this, that, these, those)
Entropy	Shannon entropy	3	6.3	4.14 (0.06)	4.12 (0.07)	0.447	See **Table 4**
FocusPast	Psychological processes	3	4	1.23 (1.43)	2.14 (2.07)	0.334	% words that are focused on the past e.g., ago, did
PosEmo	Psychological processes	3	3	2.19 (1.99)	1.19 (1.67)	0.248	% words that reflect positive emotion e.g., love, nice
S –> S CC S	Grammatical constituents	3	2.3	0.00 (0.01)	0.01 (0.01)	0.239	Two sentences joined by a coordinating conjunction e.g., "She runs but he walks."
MRC Imageability AW	Psycholinguistics	2	5.5	359.80 (13.58)	343.57 (20.67)	0.084	Mean ease of imageability of a word according to the Medical research council database. Higher values = easier imagery.
MATTR_30	Lexical richness	2	3.5	0.77 (0.04)	0.76 (0.05)	0.703	Moving average type-token-ratio with a window of 30 words

** = $p < 0.05$,** = $p < 0.001$. Features in bold appear important for classification using both overlearned narrative recall and picture description (see **Table 4**).*

validation folds for ONR samples. A further 14 features were selected only in one fold and were not considered for further analysis.

Between-group comparisons of the values of the features selected in the HC vs. AD+MCI classification using the ONR sample are displayed in **Table 4**. Seven of these features differed significantly: the mean frequency for content words measured according to the British National Corpus (BNC); Shannon entropy for letters; rate of prepositional phrases; percentage of words relating to health; number of syllables per word; the percentage of words longer than six letters; and coherence between adjacent sentences. Comparative scaled values are displayed in **Figure 1**.

Picture Description Features

Eleven features were selected in at least two folds using PD samples to classify HC vs. AD+MCI (**Table 5**). A further 11 features were selected only in one fold and eliminated from further analysis.

Group comparisons showed significant differences between the values of three features: noun phrases consisting of a bare determiner, emotional tone and sentences composed of an adverbial phrase, noun phrase and verb phrase. Comparative scaled values are plotted in **Figure 2**.

Comparisons of the selected features between the two discourse types reveal that both classifiers learned class membership from grammatical constituents, psycholinguistics and psychological processes (**Tables 4** and **5**). Two individual features (noun phrases consisting of bare determiners, and entropy) were important to both tasks. By contrast, features relating to semantic richness (Idea Density) and coherence (Mean WMD), as well as word complexity (DESWLsy and Sixltr) were important only for classification in ONR, while lexical richness (MATTR_30) was important only in PD. Moreover, a greater number of features important to the classification of ONR than to the classification of PD showed differences in values between groups.

Accuracies in Subgroup Classifications

MCI and AD subgroups were explored, as important clinically distinctive groups that may differ in management and disease course.

Healthy Controls *Versus* Alzheimer's disease

Table 6 reports classification performance for HC vs. AD alone. The highest mean balanced accuracy was achieved with ONR samples (0.90), higher than accuracy classifying the mixed AD+MCI group and balanced accuracy for the MCI alone group (both 0.78). AUC, sensitivity and specificity were also highest of all tasks (0.94, 0.83, and

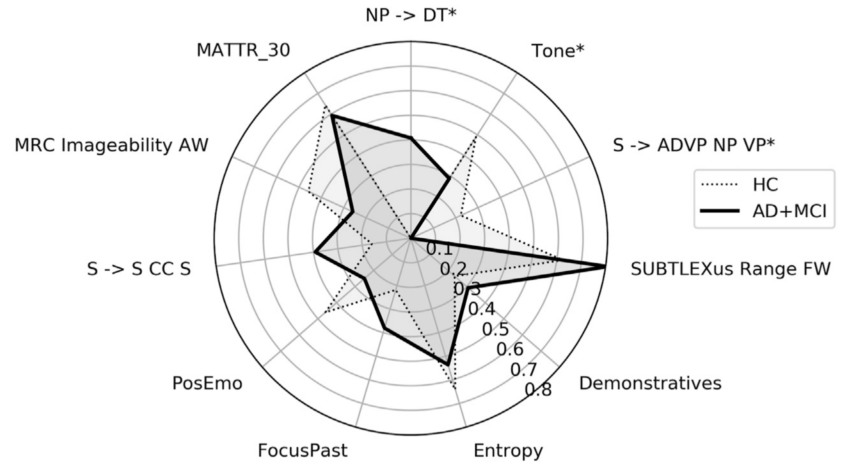


FIGURE 2 | Radar plot showing features important for HC vs. AD+MCI classification using picture description. HC = healthy control, AD+MCI = Alzheimer's disease and Mild Cognitive Impairment group. Features have been scaled to between 0 and 1 using MinMax scaling and medians plotted. * = p < 0.05.

TABLE 6 | HC vs. AD mean (s.d) SVM classification performance across five-fold cross-validation for five connected speech tasks, ranked by accuracy.

Discourse-generating task	Balanced accuracy	AUC	Sensitivity	Specificity
ONR	0.90 (0.11)	0.94 (0.06)	0.83 (0.24)	0.96 (0.09)
CS	0.75 (0.15)	0.80 (0.23)	0.62 (0.26)	0.88 (0.12)
NNR	0.71 (0.18)	0.73 (0.26)	0.65 (0.34)	0.76 (0.22)
PR	0.68 (0.24)	0.65 (0.25)	0.52 (0.46)	0.84 (0.15)
PD	0.59 (0.30)	0.75 (0.26)	0.50 (0.35)	0.68 (0.32)

TABLE 7 | HC vs. MCI mean (s.d) SVM classification performance across five-fold cross-validation for five connected speech tasks, ranked by accuracy.

Discourse-generating task	Balanced accuracy	AUC	Sensitivity	Specificity
ONR	0.78 (0.13)	0.82 (0.22)	0.67 (0.31)	0.90 (0.10)
CS	0.70 (0.20)	0.75 (0.10)	0.58 (0.37)	0.82 (0.19)
PD	0.62 (0.26)	0.77 (0.28)	0.40 (0.42)	0.84 (0.15)
PR	0.52 (0.12)	0.62 (0.21)	0.43 (0.25)	0.60 (0.19)
NNR	0.50 (0.23)	0.45 (0.30)	0.27 (0.43)	0.73 (0.18)

0.96 respectively). PD produced the lowest balanced accuracy (0.59) and sensitivity (0.50), but third highest AUC (0.75).

Healthy Controls *Versus* Mild Cognitive Impairment

Table 7 reports classification performance for HC vs. MCI alone. The pattern of speech task performance more closely resembles that of HC vs. AD+MCI (**Table 3**); ONR achieved the highest balanced accuracy (0.78), AUC (0.82), sensitivity (0.67) and specificity (0.90) and NNR produced the lowest balanced accuracy (0.50), AUC (0.50) and sensitivity (0.27). Only the two top performing tasks (ONR and CS) reached sensitivity above chance level.

Comparing the three classifications, performance was higher in all four metrics for HC vs. AD compared to HC vs. MCI, and HC vs. AD+MCI (**Figure 3**). Accuracy/balanced accuracy was equal for both HC vs. MCI and HC vs AD+MCI classifications (0.78); AUC and sensitivity were higher for HC vs. AD+MCI but

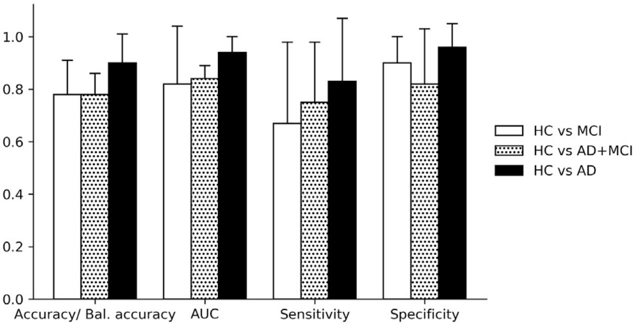


FIGURE 3 | Classification performance for groups and subgroups. HC = healthy control, MCI = Mild Cognitive Impairment, AD = Alzheimer's disease, AD+MCI = Alzheimer's disease and Mild Cognitive Impairment group. All classifications used linguistic features from the overlearned narrative recall task. Error bars + 1 sd.

specificity was lower, suggesting poorer correct classification of HC given a mixed patient group, compared to MCI only.

Demographic Variables

A linear regression with the twelve important features from ONR (**Table 4**) as input failed to predict age (whole sample $r^2 = -0.14$, HC alone $r^2 = -11.9$, AD+MCI alone $r^2 = -1.33$) or years in education (whole sample $r^2 = -0.30$, HC alone $r^2 = -11.83$, AD+MCI alone $r^2 = -5.60$)[2]. Balanced accuracy for classification of sex was greater than chance (0.55), however the male/female split included both HC and AD+MCI participants in both groups.

DISCUSSION

The accuracy of linguistic features automatically extracted from five connected speech tasks for classifying mild AD and MCI was compared. Differences were observed in classification performance using SVM, which, although small for the top performing tasks, indicated differential clinical utility for classifying mild AD and MCI based on task choice.

When comparing cognitively healthy controls with those judged likely on clinical grounds to harbor AD pathology, (i.e. diagnosed with either MCI or AD) the highest accuracy (78%) was achieved using data obtained using ONR. The same data also yielded the highest accuracy in smaller, but clinically relevant, subgroup classifications (mild AD alone or MCI alone compared to HC (90% and 78% respectively)). These results suggest that an overlearned narrative recall task may be the best approach to obtaining discourse samples for detecting early or pre-symptomatic cases of AD, a goal that has become central to successful clinical trial outcomes.

PD achieved the second highest accuracy (76%) supporting the role of a new, updated version of this commonly used task. Sensitivity was lower (69% compared to 75% for ONR), and the task performed poorly for classification of AD only. The accuracy of features probably increases with sample length (Fraser et al., 2016), so the shorter samples obtained from the AD group may have hindered classification. PR, which is also a short task, achieved the third highest accuracy (74%) and was ranked third for detecting mild AD and fourth for MCI.

Although conversational discourse elicited using a map reading task achieved only 66% accuracy to detect AD+MCI, accuracy improved in the subgroup analyses: CS gave the second highest accuracy for mild AD and MCI groups alone, suggesting that critical differences in CS may develop between the MCI and mild dementia stages.

NNR with a picture-book stimulus produced the worst performance for AD+MCI and the MCI subgroup classification. In a previous study in which retellings of the same task were scored by a linguist, only 15% of AD patients grasped the overall theme of the story (Ash et al., 2007). Fine-grained linguistic features alone are unlikely to capture this deficiency and global scoring has not yet been adequately automatized (though see Dunn et al. (2002) for a potential approach based on Latent Semantic Analysis).

The minimum sample length required for meaningful analysis has been subject to debate (Sajjadi et al., 2012). Our main results (AD+MCI classification) suggest that little accuracy is lost when classifying shorter samples (PD and PR), and the lowest accuracy was achieved using the longest samples (NNR). Conditions of the task may therefore be of more importance than resulting sample length, useful for clinical adoption. However, when little data is available, and samples are short (such as in the AD alone classification), classification performance may suffer.

Features Important for Classification

Although the advantage of ONR may simply be task-related, (i.e. due to the involvement of memory as well as language), it is also instructive to examine features that were robustly selected and the overlap with those selected from PD samples. As in Sajjadi et al. (2012) and Beltrami et al. (2016) a multi-domain linguistic impairment was detected in the patient group, with changes evident in lexical, semantic and syntactic features, and speech tasks showing varying sensitivity to these changes.

Word Frequency

In keeping with the findings of Garrard et al. (2005) and those of Masrani et al. (2017) participants in the AD+MCI group used words with higher lexical frequency. Studies of patients with isolated degradation of semantic knowledge due to focal left anterior temporal atrophy semantic dementia (SD) have found that specific terms are replaced with higher frequency generic usages (Bird et al., 2000; Fraser et al., 2014; Meteyard et al., 2014). Word frequency can therefore be seen as reflecting the integrity of the brain's store of world knowledge, a deficit that is seen in a high proportion of patients with early AD (Hodges et al., 1992).

Entropy

Entropy was retained in five folds using ONR, and three for PD. Entropy quantifies the information content contained in a string of letters (Shannon, 1951): the more predictable a letter is on the basis of those that come before it, the lower its entropy. Averaged over letters, entropy was significantly lower in the AD+MCI group using ONR, suggesting greater predictability in these samples. Entropy in discourse samples elicited using PD correlates with global cognition (Hernández-Domínguez et al., 2018), and the findings of the current study also suggest that lower values are indicative of early AD, and that this is constant across tasks. Lower levels of entropy may inherently vary between tasks Chen et al. (2017); the current study found lower values in ONR than in PD discourse, with between-group differences significant in the former. The value of entropy may therefore be greater when considering more cognitively demanding tasks.

Emotional Tone

The overall emotional tone (a "summary variable" calculated by LIWC2015 (Pennebaker et al., 2015)) of the sample was an

[2] Negative r^2 values indicate that predicting the mean dependent variable for each instance would explain more variance than a model based on the input feature.

important feature in PD, with the tone adopted by the AD+MCI group significantly more negative than HC. The same did not apply in ONR samples, for which the emotional tone is more tightly constrained by the story itself. Use of positive words was also lower in the AD+MCI group. Individuals with depression use more negative words in their writing (Rude et al., 2004), and depression commonly coexists with AD, for which it may also be a risk factor in older adults (Kitching, 2015; Herbert and Lucassen, 2016).

Grammatical Constituents

Classifications based on both ONR and PD retained in all folds the increased frequency with which participants in the AD + MCI group formed a noun phrase using a bare determiner (NP – > DT), e.g. "look at this" as opposed to "look at this jar". Determiners can serve a deictic purpose, so speech tasks with a pictorial stimulus may be more sensitive to their use; Sajjadi et al. (2012) reported a greater proportion of function words, including determiners, in PD than CS, and the difference between groups in the current study was significant for PD only. Greater numbers of determiners (Petti et al., 2020) and fewer nouns (Bucks et al., 2000; Jarrold et al., 2014) have been independently reported as features of AD discourse, but it is likely that specifying the role of the determiner in the sentence (as in NP –> DT) adds discriminatory power. A similar interpretation may obtain in the case of sentences consisting of an adverbial phrase, noun phrase and verb phrase (S –> ADVP NP V), which were also more frequent in HC discourse and may either denote richer descriptions of the picture, or a greater tendency to relate utterances to one another, e.g. by using "then".

Remaining Features

We make note of two remaining features: imageability (MRC Imageability AW) and word-movers distance (WMD). Although selected in fewer than five folds, median imageability measured in PD was numerically lower in the AD+MCI group. This "reverse imageability effect" has also been observed in speech of SD patients (Bird et al., 2000; Hoffman et al., 2014), and can be explained as a consequence of reliance on a more generic, and thus higher frequency, vocabulary: consider the less imageable "place" and the more imageable "cathedral" (Bird et al., 2000; Hoffman et al., 2014).

The mean WMD, although retained in only two folds of the ONR classifier, was significantly different between groups. Using word2vec embeddings, WMD measures the minimum cumulative distance required to travel between collections of word vectors in a high-dimensional semantic space, analogous with coherence (Mikolov et al., 2013; Kusner et al., 2015). Other measures of coherence, however, are based on the cosine of the angle between the vectors of consecutive sentences, which requires multiple word vectors to be combined into a sentence vector (Dunn et al., 2002; Holshausen et al., 2014; Mirheidari et al., 2018). This step is obviated by WMD. To the best of the author's knowledge this is the first study to show WMD as a discriminatory feature of AD and MCI speech. The measure may show differences in ONR alone because the presence of the stimulus in PD acts as a continuous referential prompt, facilitating the coherent connection of sequential utterances.

Strengths and Limitations

Demographic variables were not balanced across groups, unfortunately a common issue (de la Fuente Garcia et al., 2020). Given that the linguistic function of participants pre-diagnosis is not known, conclusions regarding between-group differences are drawn with caution. We have explored demographic variables and find little evidence of mediation, although they may still act as moderators. The population studied is small, which may account for small differences in accuracy observed for the three highest scoring tasks classifying AD+MCI. Subgroup sizes are further reduced, and these results are therefore less reliable. We have attempted to improve reliability by reporting results of cross-validation. Hyper-parameters were not tuned, e.g. via a grid search, which may improve results.

Acoustic features were not studied as extraction was beyond the scope of the study—their inclusion may have improved performance, seen in previous research such as Fraser et al. (2016) and Beltrami et al. (2016). One strength is that our AD+MCI group (and AD subgroup) were more mildly affected than those classified in Fraser et al. (2016) (mean MMSE 18.5, compared to AD+MCI mean of 24 and AD subgroup mean of 22.5), and so likely represent a more challenging classification task.

Compared to current tests, the reported AUC for detecting MCI is higher than the MMSE (82% compared to 74% (Ciesielska et al., 2016)) with similar sensitivity but better specificity (67% and 90% compared to 66% and 73%). Compared to FDG-PET for AD detection sensitivity is slightly lower with better specificity (83% and 96% compared to 86% for both (Patwardhan et al., 2004)).

Conclusion and Future Work

The results of the current study indicate that linguistic analysis could be used to detect mild AD and MCI, as well as these subgroups compared to healthy controls - an important clinical task – in a novel dataset. Computational analysis of language would offer a rapid, scalable and low-cost assessment of individuals, that could be built in to remote assessment, such as via a smartphone app, less obtrusive and anxiety provoking than current biomarker tests. We have shown, in a direct comparison of the same participants, that the choice of speech task impacts subsequent performance of classifiers trained to recognize mild AD and MCI based on linguistic features. Tasks that probe memory and language may be optimal. Although some features appear important for classification independent of discourse type, tasks may be sensitive to different linguistic features in early AD; due to the reliance on PD in previous studies, some features susceptible to disease may have garnered less attention. This has implications for future work seeking to characterize AD and MCI based on speech, and clinical adoption of computational approaches. Future work could look to explore use of different tasks in larger samples, and include novel features found here important in classifying groups to improve sensitivity to disease, such as the WMD and analysis of emotional tone.

Longitudinal assessment of healthy individuals prior to a possible later diagnosis of AD is needed, in order to identify very early linguistic changes and delineate the impact of Alzheimer pathology on language from other factors. Such studies are underway and beginning to provide important insights (Mueller et al., 2018).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Service Committee London-Dulwich. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to conception and design of the study. NC collected the data, performed the analysis and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was funded by the Medical Research Council (grant number MR/N013638/1).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2021.634360/full#supplementary-material

REFERENCES

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected Speech as a Marker of Disease Progression in Autopsy-Proven Alzheimer's Disease. *Brain* 136 (12), 3727–3737. doi:10.1093/brain/awt269

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (2011). The Diagnosis of Mild Cognitive Impairment Due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease. *Alzheimer's Demen.* 7 (3), 270–279. doi:10.1016/j.jalz.2011.03.008

Asgari, M., Kaye, J., and Dodge, H. (2017). Predicting Mild Cognitive Impairment from Spontaneous Spoken Utterances. *Alzheimer's Demen. Translational Res. Clin. Interventions* 3 (2), 219–228. doi:10.1016/j.trci.2017.01.006

Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., and Grossman, M. (2006). Trying to Tell a Tale: Discourse Impairments in Progressive Aphasia and Frontotemporal Dementia. *Neurology* 66 (9), 1405–1413. doi:10.1212/01.wnl.0000210435.72614.38

Ash, S., Moore, P., Vesely, L., and Grossman, M. (2007). The Decline of Narrative Discourse in Alzheimer's Disease. *Brain Lang.* 103 (1), 181–182. doi:10.1016/j.bandl.2007.07.105

Beltrami, D., Calzà, L., Gagliardi, G., Ghidoni, E., Marcello, N., Favretti, R. R., et al. (2016). "Automatic Identification of Mild Cognitive Impairment through the Analysis of Italian Spontaneous Speech Productions," in LREC, Portorož, Slovenia, May 23–28, 2016, 16, 2086–2093.

Berisha, V., Wang, S., LaCross, A., and Liss, J. (2015). Tracking Discourse Complexity Preceding Alzheimer's Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *Jad* 45 (3), 959–963. doi:10.3233/jad-142763

Berube, S., Nonnemacher, J., Demsky, C., Glenn, S., Saxena, S., Wright, A., et al. (2019). Stealing Cookies in the Twenty-First Century: Measures of Spoken Narrative in Healthy versus Speakers with Aphasia. *Am. J. Speech Lang. Pathol.* 28 (1S), 321–329. doi:10.1044/2018_AJSLP-17-0131

Bird, H., Lambon Ralph, M. A., Patterson, K., and Hodges, J. R. (2000). The Rise and Fall of Frequency and Imageability: Noun and Verb Production in Semantic Dementia. *Brain Lang.* 73 (1), 17–49. doi:10.1006/brln.2000.2293

Bloudek, L. M., Spackman, D. E., Blankenburg, M., and Sullivan, S. D. (2011). Review and Meta-Analysis of Biomarkers and Diagnostic Imaging in Alzheimer's Disease. *Jad* 26 (4), 627–645. doi:10.3233/jad-2011-110458

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., and Cappa, S. F. (2017). Connected Speech in Neurodegenerative Language Disorders: A Review. *Front. Psychol.* 8, 269. doi:10.3389/fpsyg.2017.00269

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The Balanced Accuracy and its Posterior Distribution. *Proc.—Int. Conf. Pattern Recognition*, 3121–3124. doi:10.1109/ICPR.2010.764

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of Spontaneous, Conversational Speech in Dementia of Alzheimer Type: Evaluation of an Objective Technique for Analysing Lexical Performance. *Aphasiology* 14 (1), 71–91. doi:10.1080/026870300401603

Carnero-Pardo, C. (2014). Should the Mini-Mental State Examination Be Retired?. *Neurología (English Edition)* 29 (8), 473–481. doi:10.1016/j.nrleng.2013.07.005

Chen, R., Liu, H., and Altmann, G. (2017). Entropy in Different Text Types. *Digital Scholarship Humanities* 32 (3), fqw008–542. doi:10.1093/llc/fqw008

Ciesielska, N., Sokołowski, R., Mazur, E., Podhorecka, M., Polak-Szabela, A., and Kędziora-Kornatowska, K. (2016). Is the Montreal Cognitive Assessment (MoCA) Test Better Suited Than the Mini-Mental State Examination (MMSE) in Mild Cognitive Impairment (MCI) Detection Among People Aged over 60? Meta-Analysis. *Psychiatr. Pol.* 50 (5), 1039–1052. doi:10.12740/pp/45368

Clarke, N., Foltz, P., and Garrard, P. (2020). How to Do Things with (Thousands of) Words: Computational Approaches to Discourse Analysis in Alzheimer's Disease. *Cortex* 129, 446–463. doi:10.1016/j.cortex.2020.05.001

Cummings, J., Aisen, P. S., Dubois, B., Frölich, L., Jack, C. R., Jones, R. W., et al. (2016). Drug Development in Alzheimer's Disease: the Path to 2025. *Alz Res. Ther.* 8 (1), 1–12. doi:10.1186/s13195-016-0207-9

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *J. Alzheimers Dis.*, (Preprint), 1–27.

Drummond, C. u., Coutinho, G., Fonseca, R. P., AssunÃ§Ã£o, N., Teldeschi, A., de Oliveira-Souza, R., et al. (2015). Deficits in Narrative Discourse Elicited by Visual Stimuli Are Already Present in Patients with Mild Cognitive Impairment. *Front. Aging Neurosci.* 7. doi:10.3389/fnagi.2015.00096

Dunn, J. C., Almeida, O. P., Barclay, L., Waterreus, A., and Flicker, L. (2002). Latent Semantic Analysis: A New Method to Measure Prose Recall. *J. Clin. Exp. Neuropsychol.* 24 (1), 26–35. doi:10.1076/jcen.24.1.26.965

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental State". *J. Psychiatr. Res.* 12, 189–198. doi:10.1016/0022-3956(75)90026-6

Forbes-McKay, K. E., and Venneri, A. (2005). Detecting Subtle Spontaneous Language Decline in Early Alzheimer's Disease with a Picture Description Task. *Neurol. Sci.* 26 (4), 243–254. doi:10.1007/s10072-005-0467-9

Fraser, K. C., Lundholm Fors, K., Eckerström, M., Öhman, F., and Kokkinakis, D. (2019). Predicting MCI Status from Multimodal Language Data Using Cascaded Classifiers. *Front. Aging Neurosci.* 11, 205. doi:10.3389/fnagi.2019.00205

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., et al. (2014). Automated Classification of Primary Progressive Aphasia Subtypes from Narrative Speech Transcripts. *Cortex* 55 (1), 43–60. doi:10.1016/j.cortex.2012.12.006

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2015). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Jad* 49, 407–422. doi:10.3233/JAD-150520

Garrard, P. (2009). Cognitive Archaeology: Uses, Methods, and Results. *J. Neurolinguist.* 22 (3), 250–265. doi:10.1016/j.jneuroling.2008.07.006

Garrard, P., Haigh, A.-M., and de Jager, C. (2011). Techniques for Transcribers: Assessing and Improving Consistency in Transcripts of Spoken Language. *Literary Linguistic Comput.* 26 (4), 389–405. doi:10.1093/llc/fqr018

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2005). The Effects of Very Early Alzheimer's Disease on the Characteristics of Writing by a Renowned Author. *Brain* 128 (2), 250–260. doi:10.1093/brain/awh341

Goodglass, H., Kaplan, E., and Weintraub, S. (1983). Boston Naming Test. *Lea and Febiger*.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learn.* 46, 389–422. doi:10.1007/978-3-540-88192-6-810.1023/a:1012487302797

Herbert, J., and Lucassen, P. J. (2016). Depression as a Risk Factor for Alzheimer's Disease: Genes, Steroids, Cytokines and Neurogenesis—what Do We Need to Know?. *Front. Neuroendocrinology* 41, 153–171. doi:10.1016/j.yfrne.2015.12.001

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., and Roche-Bergua, A. (2018). Computer-based Evaluation of Alzheimer's Disease and Mild Cognitive Impairment Patients during a Picture Description Task. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 10, 260–268. doi:10.1016/j.dadm.2018.02.004

Hodges, J. R., Salmon, D. P., and Butters, N. (1992). Semantic Memory Impairment in Alzheimer's Disease: Failure of Access or Degraded Knowledge?. *Neuropsychologia* 30 (4), 301–314. doi:10.1016/0028-3932(92)90104-t

Hoffman, P., Meteyard, L., and Patterson, K. (2014). Broadly Speaking: Vocabulary in Semantic Dementia Shifts towards General, Semantically Diverse Words. *Cortex* 55 (1), 30–42. doi:10.1016/j.cortex.2012.11.004

Holshausen, K., Harvey, P. D., Elvevåg, B., Foltz, P. W., and Bowie, C. R. (2014). Latent Semantic Variables Are Associated with Formal Thought Disorder and Adaptive Behavior in Older Inpatients with Schizophrenia. *Cortex* 55 (1), 88–96. doi:10.1016/j.cortex.2013.02.006

Hsieh, S., Schubert, S., Hoon, C., Mioshi, E., and Hodges, J. R. (2013). Validation of the Addenbrooke's Cognitive Examination III in Frontotemporal Dementia and Alzheimer's Disease. *Dement Geriatr. Cogn. Disord.* 36, 242–250. doi:10.1159/000351671

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical Model of Dynamic Biomarkers of the Alzheimer's Pathological Cascade. *Lancet Neurol.* 9 (1), 119–128. doi:10.1016/S1474-4422(09)70299-6

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., et al. (2014). Aided Diagnosis of Dementia Type through Computer-Based Analysis of Spontaneous Speech. *Proc. Workshop Comput. Linguistics Clin. Psychol. Linguistic Signal Clin. Reality*, 27–37. doi:10.3115/v1/W14-3204

Kitching, D. (2015). Depression in Dementia. *Aust. Prescr* 38 (6), 209–211. doi:10.18773/austprescr.2015.071

Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. (2015). "From Word Embeddings to Document Distances," in Proceedings of the 32nd International Conference on Machine Learning, 957–966.

Laske, C., Sohrabi, H. R., Frost, S. M., López-de-Ipiña, K., Garrard, P., Buscema, M., et al. (2015). Innovative Diagnostic Tools for Early Detection of Alzheimer's Disease. *Alzheimer's Demen.*, 11, 561–578. doi:10.1016/j.jalz.2014.06.004

Lombardi, G., Crescioli, G., Cavedo, E., Lucenteforte, E., Casazza, G., Bellatorre, A. G., and Filippini, G. (2020). Structural Magnetic Resonance Imaging for the Early Diagnosis of Dementia Due to Alzheimer's Disease in People with Mild Cognitive Impairment. *Cochrane Database Syst. Rev.* (3). doi:10.1002/14651858.cd009628.pub2

Lovestone, S. (2014). Blood Biomarkers for Alzheimer's Disease. *Genome Med.* 6 (8), 8–11. doi:10.1186/s13073-014-0065-7

MacWhinney, B. (2019). Understanding Spoken Language through TalkBank. *Behav. Res.* 51 (4), 1919–1927. doi:10.3758/s13428-018-1174-9

Masrani, V., Murray, G., Field, T., and Carenini, G. (2017). "Detecting Dementia through Retrospective Analysis of Routine Blog Posts by Bloggers with Dementia," in *BioNLP 2017*, 232–237. Retrieved from:http://www.aclweb.org/anthology/W17-2329 (Accessed September 12, 2017).

Matias-Guiu, J. A., Cortés-Martínez, A., Valles-Salgado, M., Rognoni, T., Fernández-Matarrubia, M., Moreno-Ramos, T., et al. (2016). Addenbrooke's Cognitive Examination III: Diagnostic Utility for Mild Cognitive Impairment and Dementia and Correlation with Standardized Neuropsychological Tests. *Int. Psychogeriatr.* 29 (1), 105–113. doi:10.1017/S1041610216001496

Matías-Guiu, J. A., Pytel, V., Cortés-Martínez, A., Valles-Salgado, M., Rognoni, T., Moreno-Ramos, T., et al. (2018). Conversion between Addenbrooke's Cognitive Examination III and Mini-Mental State Examination. *Int. Psychogeriatr.* 30 (8), 1227–1233. doi:10.1017/S104161021700268X

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., et al. (2011). The Diagnosis of Dementia Due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease. *Alzheimer's Demen.* 7 (3), 263–269. doi:10.1016/j.jalz.2011.03.005

Meteyard, L., Quain, E., and Patterson, K. (2014). Ever Decreasing Circles: Speech Production in Semantic Dementia. *Cortex* 55 (1), 17–29. doi:10.1016/j.cortex.2013.02.013

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, 1–12. doi:10.1162/153244303322533223

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019). Dementia Detection Using Automatic Analysis of Conversations. *Computer Speech Lang.* 53, 65–79. doi:10.1016/j.csl.2018.07.006

Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2018). Detecting Signs of Dementia Using Word Vector Representations. Proc. Interspeech 2018, Hyderabad, India, September 2–6, 2018, 1893–1897. doi:10.21437/Interspeech.2018-1764

Mitchell, A. J. (2009). CSF Phosphorylated Tau in the Diagnosis and Prognosis of Mild Cognitive Impairment and Alzheimer's Disease: a Meta-Analysis of 51 Studies. *J. Neurol. Neurosurg. Psychiatry* 80 (9), 966–975. doi:10.1136/jnnp.2008.167791

Mueller, K. D., Koscik, R. L., Hermann, B. P., Johnson, S. C., and Turkstra, L. S. (2018). Declines in Connected Language Are Associated with Very Early Mild Cognitive Impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Front. Aging Neurosci.* 9, 1–14. doi:10.3389/fnagi.2017.00437

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool for Mild Cognitive Impairment. *J. Am. Geriatr. Soc.* (53), 695–699. doi:10.1029/WR017i002p00410

Navarro, G. (2001). A Guided Tour to Approximate String Matching. *ACM Comput. Surv.* 33 (1), 31–88. doi:10.1145/375360.375365

Orimaye, S. O., Wong, J. S.-M., and Wong, C. P. (2018). Deep Language Space Neural Network for Classifying Mild Cognitive Impairment and Alzheimer-type Dementia. *PLoS ONE* 13 (11), e0205636–15. doi:10.1371/journal.pone.0205636

Patwardhan, M. B., McCrory, D. C., Matchar, D. B., Samsa, G. P., and Rutschmann, O. T. (2004). Alzheimer Disease: Operating Characteristics of PET- A Meta-Analysis. *Radiology* 231 (1), 73–80. doi:10.1148/radiol.2311021620

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Petersen, R. C. (2004). Mild Cognitive Impairment as a Diagnostic Entity. *J. Intern. Med.* 256, 183–194. doi:10.1111/j.1365-2796.2004.01388.x

Petti, U., Baker, S., and Korhonen, A. (2020). A Systematic Literature Review of Automatic Alzheimer's Disease Detection from Speech and Language. *J. Am. Med. Inform. Assoc.* 27 (0), 1784–1797. doi:10.1093/jamia/ocaa174

Rude, S., Gortner, E.-M., and Pennebaker, J. (2004). Language Use of Depressed and Depression-Vulnerable College Students. *Cogn. Emot.* 18 (8), 1121–1133. doi:10.1080/02699930441000030

Sajjadi, S. A., Patterson, K., Tomek, M., and Nestor, P. J. (2012). Abnormalities of Connected Speech in Semantic Dementia v.s Alzheimer's Disease. *Aphasiology* 26 (6), 847–866. doi:10.1080/02687038.2012.654933

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell Syst. Tech. J.* 30 (1), 50–64. doi:10.1002/j.1538-7305.1951.tb01366.x

Sherratt, S., and Bryan, K. (2019). Textual Cohesion in Oral Narrative and Procedural Discourse: the Effects of Ageing and Cognitive Skills. *Int. J. Lang. Commun. Disord.* 54 (1), 95–109. doi:10.1111/1460-6984.12434

Thompson, H. S., Anderson, A., Bard, E. G., Doherty-Sneddon, G., Newlands, A., and Sotillo, C. (1993). The HCRC Map Task Corpus. *Proc. Workshop Hum. Lang. Technology*, 25–30. doi:10.3115/1075671.1075677

Toledo, C. M., Aluísio, S. M., Santos, L. B., Brucki, S. M. D., Trés, E. S., Oliveira, M. O., et al. (2017). Analysis of Macrolinguistic Aspects of Narratives from Individuals with Alzheimer's Disease, Mild Cognitive Impairment, and No Cognitive Impairment. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 10, 31–40. doi:10.1016/j.dadm.2017.08.005

Conflict of Interest: The 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.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 14 June 2021
doi: 10.3389/fnagi.2021.642647

Temporal Integration of Text Transcripts and Acoustic Features for Alzheimer's Diagnosis Based on Spontaneous Speech

Matej Martinc[1]*, *Fasih Haider*[2†], *Senja Pollak*[1†] *and Saturnino Luz*[2*†]

[1] *Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia,* [2] *Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh, United Kingdom*

Background: Advances in machine learning (ML) technology have opened new avenues for detection and monitoring of cognitive decline. In this study, a multimodal approach to Alzheimer's dementia detection based on the patient's spontaneous speech is presented. This approach was tested on a standard, publicly available Alzheimer's speech dataset for comparability. The data comprise voice samples from 156 participants (1:1 ratio of Alzheimer's to control), matched by age and gender.

Materials and Methods: A recently developed Active Data Representation (ADR) technique for voice processing was employed as a framework for fusion of acoustic and textual features at sentence and word level. Temporal aspects of textual features were investigated in conjunction with acoustic features in order to shed light on the temporal interplay between paralinguistic (acoustic) and linguistic (textual) aspects of Alzheimer's speech. Combinations between several configurations of ADR features and more traditional bag-of-n-grams approaches were used in an ensemble of classifiers built and evaluated on a standardised dataset containing recorded speech of scene descriptions and textual transcripts.

Results: Employing only semantic bag-of-n-grams features, an accuracy of 89.58% was achieved in distinguishing between Alzheimer's patients and healthy controls. Adding temporal and structural information by combining bag-of-n-grams features with ADR audio/textual features, the accuracy could be improved to 91.67% on the test set. An accuracy of 93.75% was achieved through late fusion of the three best feature configurations, which corresponds to a 4.7% improvement over the best result reported in the literature for this dataset.

Conclusion: The proposed combination of ADR audio and textual features is capable of successfully modelling temporal aspects of the data. The machine learning approach toward dementia detection achieves best performance when ADR features are combined with strong semantic bag-of-n-grams features. This combination leads to state-of-the-art performance on the AD classification task.

Keywords: Alzheimer's dementia detection, speech, language, acoustic features, lexical features, natural language processing, speech processing, machine learning

OPEN ACCESS

Edited by:
Annalena Venneri,
The University of Sheffield,
United Kingdom

Reviewed by:
Ronan O'Malley,
The University of Sheffield,
United Kingdom
Daniel Blackburn,
Sheffield Teaching Hospitals NHS Foundation Trust, United Kingdom
Heidi Christensen,
The University of Sheffield,
United Kingdom

***Correspondence:**
Saturnino Luz
s.luz@ed.ac.uk
Matej Martinc
matej.martinc@ijs.si

†ORCID:
Fasih Haider
orcid.org/0000-0002-5150-3359
Senja Pollak
orcid.org/0000-0002-4380-0863
Saturnino Luz
orcid.org/0000-0001-8430-7875

Received: *16 December 2020*
Accepted: *11 May 2021*
Published: *14 June 2021*

Citation:
Martinc M, Haider F, Pollak S and Luz S (2021) Temporal Integration of Text Transcripts and Acoustic Features for Alzheimer's Diagnosis Based on Spontaneous Speech. Front. Aging Neurosci. 13:642647. doi: 10.3389/fnagi.2021.642647

1. INTRODUCTION

While the natural history of Alzheimer's Disease (AD) and the form of dementia it causes are mainly characterised by memory impairment, a wide range of cognitive functions are known to be affected by the process of neurodegeneration triggered by the disease. Several standardised neuropsychological tests are currently employed to detect such impairments for the purposes of diagnosis and assessment of disease progression. However, these tests often take place in clinics and consist of constrained cognitive tasks, where the patient's performance may be affected by extraneous factors such as variations in mood, poor sleep the night before the test, etc. Recent progress in artificial intelligence (AI) and machine learning (ML) technology has opened new avenues for more comprehensive monitoring of cognitive function, and tests based on spontaneous speech and language data have emerged as possible tools for diagnostic and prognostic assessment (de la Fuente Garcia et al., 2020; Petti et al., 2020). In this paper we investigate the hypothesis that integration of acoustic and textual data into a unified ML model enhances the accuracy of AD detection. Specifically, we present a model that integrates acoustic and textual modalities on a temporal (i.e., time-based) dimension. The motivation for doing so arises from the nature of the task used to elicit the speech data used in this study. These data consist of spontaneous speech elicited through the Cookie Theft description task from the Boston Diagnostic Aphasia Exam (Goodglass et al., 2001), which involves visuospatial as well as verbal ability.

Along with language, visuospatial function is affected early in AD. This is manifested in the form of non-salience of visual input stimulus, and degraded attentional focus and visual search, among other disturbances (Cronin-Golomb, 2011). Using a similar picture description task, Meguro et al. (2001) observed hemispatial visual searching impairment in some participants with AD, in correlation with decreased contralateral parietal blood flow. Other studies involving picture descriptions have associated AD with simultanagnosia, a disorder of attentive exploration of the spatial field (Vighetto, 2013). They found that persons with AD tended to produce "slow and partial [descriptions], one detail after the other, without ability to capture a global perception of the drawing." Our assumption is that such disturbances of visuospatial function will be reflected in differences in temporal order between the descriptions produced by participants with AD and those produced by non-AD participants. As Cummings (2019) observed, while the Cookie Theft picture is a static scene, causal and temporal relations can be inferred from the various elements depicted in it. Capturing these relations is necessary to give a complete description of the picture, as "certain events in the scene must take place before other events in order for a description of the picture to make sense." If, as seems likely, degraded attention focus hinders the participant's ability to identify such events, one should expect the temporal organisation of events in the scene description to differ in AD.

We therefore propose an approach to speech and language which incorporates temporal information. Unlike most other approaches, where content is represented as order-agnostic features with at most short distance dependencies, our model accounts for temporal aspects of both linguistic and acoustic features. We employ our recently developed Active Data Representation (ADR) processing technique (Haider et al., 2020) and present a novel way of fusing acoustic and text features at sentence and word level. We show that these features are capable of modelling temporal aspects of text and audio, but fall short of semantic modelling. To address this shortcoming, we propose combining ADR features with term frequency-inverse document frequency weighted bag-of-n-grams features, which proved effective in modelling semantics in previous studies (Martinc and Pollak, 2020). The final combination of ADR and bag-of-n-grams features leads to state-of-the-art performance on the AD classification task[1].

2. RELATED WORK

The complex multimodal ways in which AD symptoms may appear calls for increasingly interdisciplinary research (Turner et al., 2020). Current research on AD involves not only biomedicine, neuroscience, and cognitive psychology, but also increasingly AI and machine learning methods. Studies connecting language and AD have focused mostly on formal aspects of language (i.e., lexicon, syntax and semantics), but the analysis of continuous speech has been progressively seen by researchers as a source of information that may support diagnosis of dementia and related conditions (Lopez-de Ipiña et al., 2015, 2016; Luz et al., 2018; Toth et al., 2018; Haulcy and Glass, 2021; Mahajan and Baths, 2021).

Language research into AD has employed high-level features such as information content, comprehension of complexity, picture naming and word-list generation as predictors of disease progression (Reilly et al., 2010). A study by Roark et al. (2011) used natural language processing (NLP) and automatic speech recognition (ASR) to automatically annotate and time-align a few spoken language features (pause frequency and duration), and compared these methods to manual analysis. They analysed audio recordings of 74 neuropsychological assessments to classify mild cognitive impairment (MCI) and healthy elderly participants. Their best classifier obtained an area under the receiver operating curve (AUC) of 86% by including a combination of automated speech and language features and cognitive tests scores. Jarrold et al. (2014) worked with a dataset consisting of semi-structured interviews from 9 healthy participants, 9 with AD, 9 with frontotemporal dementia, 13 with semantic dementia, and 8 with progressive nonfluent aphasia. With an automatic speech recognition (ASR) system, they extracted 41 features, including speech rate, and the mean and standard deviation of the duration of pauses, vowels, and consonants. They used a multilayered perceptron network, achieving an accuracy of 88% for AD vs. healthy subjects based on lexical and acoustic features. A more recent study by Luz et al. (2018) extracted graph-based features encoding

[1] The source code for the experiments and methods described in this paper is available under the terms of the MIT free software license at https://github.com/matejMartinc/alzheimer_diagnosis.

turn-taking patterns and speech rate (Luz, 2009) from the Carolina Conversations Collection (Pope and Davis, 2011) of spontaneous interviews of AD patients and healthy controls. Their additive logistic regression model obtained 85% accuracy in distinguishing dialogues involving an AD speaker from controls.

More recently, multimodal representations have been explored, combining linguistic and paralinguistic aspects of communication (Haider et al., 2020; Mahajan and Baths, 2021), as well as eye-tracking and other sensor modalities (Jonell et al., 2021). Those studies combined signal processing and machine learning to detect subtle acoustic signs of neurodegeneration which may be imperceptible to human diagnosticians. Toth et al. (2018), for instance, found that filled pauses (sounds like "hmmm," etc.) could not be reliably detected by human annotators, and that detection improved by using ASR-generated transcriptions. Using ASR features with a random forest classifier, Toth et al. (2018) reported an improvement over manually generated features (75 vs. 69.1% accuracy) for AD detection. Similar machine learning methods were used by König et al. (2015), who reported an accuracy of 79% when distinguishing MCI participants from healthy controls; 94% for AD vs. healthy; and 80% for MCI vs. AD. However, their tests involved different data collection procedures, including semantic fluency and sentence repetition tasks, in addition to a picture description task, with most features extracted from non-spontaneous, non-connected speech data. Motivated by the prospect of comprehensive cognitive status monitoring (Luz, 2017), studies in this field have moved toward analysis of spontaneous speech, and toward languages other than English. Weiner et al. (2016) analysed semi-structured German dialogues employing linear discriminant analysis to classify participants as healthy controls, Alzheimer's or age-associated cognitive decline, obtaining a mean accuracy score of 85.7%. This work has later been extended for prediction of development of dementia within 5 and 12 years in participants of the Interdisciplinary Longitudinal Study on Adult Development And Aging (ILSE), using a combination of acoustic and linguistic features (Weiner et al., 2019). Others have investigated the use of virtual agents as a data collection strategy for AD detection. Tanaka et al. (2017) collected dialogue, eye-tracking and video data from 29 Japanese participants who conducted structured dialogues with a virtual agent. They obtained 83% accuracy in classifying AD and control participants, using combined acoustic and textual modalities on a support vector machine (SVM) classifier. Mirheidari et al. (2019a) compared the accuracy of automated conversational analysis (ML with a combination of acoustic and linguistic features) for detection of AD on recorded doctor-patient consultations and on dialogues recorded through human-robot interaction. They reported similar accuracy for both settings using manual transcriptions ($\approx$ 90%), suggesting that automated dialogue collection could be useful in mental health monitoring.

These studies evidence the heterogeneity with which language and speech impairments are displayed in AD and related diseases. Duong et al. (2005) ran a cluster analysis with data from picture narratives and concluded that, rather than a common profile, there were several discourse patterns that could be indicative of differences between healthy ageing and AD. This heterogeneity seems to be more evident in AD than in specific disorders such as primary progressive aphasia (Ahmed et al., 2013), especially in early stages of AD (Hodges and Patterson, 1995). Therefore, we hypothesise that a comprehensive analysis of state-of-the-art paralinguistic feature sets which have been successfully used in different prediction tasks may help identify such patterns and enhance accuracy of early AD detection.

The Pitt Corpus (Becker et al., 1994), which forms part of the DementiaBank (MacWhinney, 2019), and more specifically its Cookie Theft test sub-corpus, remains one of the very few available datasets to link spontaneous speech from dementia patients and healthy controls (recordings and transcriptions) with clinical information. Therefore, this dataset has been used in several studies, including the studies by Fraser et al. (2016), Hernández-Domínguez et al. (2018), and others (Yancheva and Rudzicz, 2016; Luz, 2017; Orimaye et al., 2017; Guo et al., 2019; Mirheidari et al., 2019b; Haider et al., 2020). These studies used different combinations of information coverage measures, linguistic features and acoustic features for automatic classification of dementia under different representation methods, ranging from simple descriptive statistics to more complex feature embedding representations. Among these studies, only Mirheidari et al. (2019b) investigated the possible relation, which we discussed above, between the temporal organisation of picture descriptions and cognitive impairment. In that work, verbal references were used to simulate the participants gaze and extract features corresponding to "areas of interest." By combining such features with timing and pause information, and GloVe word vectors (Pennington et al., 2014) they were able to achieve 80% F_1 score on manually transcribed data, and $F_1 = 72\%$ on ASR outputs.

Speech research aiming at dementia detection is heterogeneous and comparisons are difficult to draw. Heterogeneity of dataset hinders comparison among the various studies on spontaneous speech for AD detection. The ADReSS challenge dataset (Luz et al., 2020) was created to mitigate this problem. In the shared task posed by ADReSS, all participants used the same dataset, which was balanced for age and gender and acoustically normalised. This is the dataset used in the present study. The various approaches proposed to tackle the ADReSS challenge included state-of-the-art deep learning and word embedding methods, and focused mainly on linguistic features extracted from the manually generated transcripts. The winning team (Yuan et al., 2020) leveraged audio recordings to obtain information about pauses in speech, encoding them as punctuation. The modified transcripts with encoded pauses were fed into an ensemble of 50 BERT (Devlin et al., 2019) and 50 ERNIE (Zhang et al., 2019) models, and majority voting was employed to derive the final predictions on the test set. They reported best accuracy (89.58%) when an ensemble of 50 ERNIE models was applied.

3. DATASET

This study uses the ADReSS subset of the Pitt Corpus, derived from a dataset gathered longitudinally between 1983 and 1988 on a yearly basis as part of the Alzheimer Research Program at the University of Pittsburgh (Becker et al., 1994; Corey Bloom and Fleisher, 2000), and made available through DementiaBank (MacWhinney, 2019). Participants are categorised into three groups: dementia, control (non-AD), and unknown status. All participants were required to be above 44 years of age, have at least 7 years of education, have no history of nervous system disorders or be taking neuroleptic medication, have an initial Mini-Mental State Examination (MMSE) score of 10 or more and be able to provide informed consent. Extensive neuropsychological and physical assessments conducted on the participants are also included (Becker et al., 1994).

While the Pitt Corpus contains data elicited through several tasks, our selected subset exclusively used the Cookie Theft description task subset, where participants were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Becker et al., 1994; Goodglass et al., 2001). This study specifically uses a subset of AD and control data matched for age and gender provided by the ADReSS challenge (Luz et al., 2020) to avoid bias, guarantee repeatability, and allow direct comparison with other ML approaches. In the following section, we provide a brief description of the methods used in the generation of the ADReSS dataset. The dataset and baseline results for the AD detection challenge are presented in detail by Luz et al. (2020).

3.1. The ADReSS Dataset

The pipeline employed in the preprocessing of the audio files is shown on the top part of **Supplementary Figure 1**. Initially, noise was sampled from short intervals from each audio recording, and subsequently spectral subtraction was applied to eliminate any noise matching those samples. Other non-target sounds such as background talk, ambulance sirens and door slamming, were minimised through selection of audio files with signal-to-noise ratio (SNR) ≥ -17 dB. Where multiple audio files existed per participant, the ADReSS organisers chose a subset that maximised audio quality and the number of samples in the matched dataset by selecting the latest recording, subjected to age and gender matching constraints. This resulted in a selection of 62 ($\approx$ 40%) recordings taken on baseline visits, 57 ($\approx$ 37%) on first visits, 19 ($\approx$ 12%) on second visits, 17 ($\approx$ 11%) on third visits and one ($<$ 1%) on the fourth visit.

As age and gender are considered major risk factors for dementia (Dukart et al., 2011), these variables are possible confounders between the AD and non-AD groups. To eliminate this possible confounding, these groups are matched for age and gender in the ADReSS dataset. For age, 5-year ranges were chosen empirically to optimise the number of recordings included in the final dataset. As a result, 156 participants matched the inclusion criteria. Of these, 78 were healthy and 78 were diagnosed with probable AD. **Supplementary Table 1** presents the demographics of the data used for training and testing. We note that the only patient in the $[50, 55)$ age interval in the AD training set had an MMSE of 30, which would not normally match the diagnosis criterion for AD. Upon detailed inspection of the Pitt metadata we found that this patient in fact had an MCI diagnosis (memory only) and therefore should not have been included in the dataset. However, we decided to keep this data point in our training set for comparability with other models trained on the ADReSS dataset.

4. TEMPORAL ANALYSIS

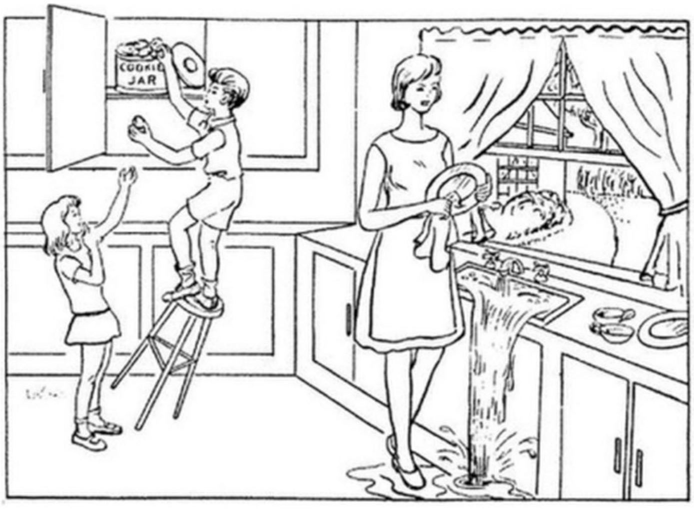
As discussed in section 1, temporal aspects of the descriptions might provide important predictors in distinguishing between AD and non-AD speech. In this section, we present a temporal analysis of the transcripts, investigating the underlying assumption that the order in which specific situations in the cookie theft picture (see **Figure 1**) are described differs. More specifically, we investigate if there is enough information available for the models to detect the temporal discrepancies between the two diagnosis groups.

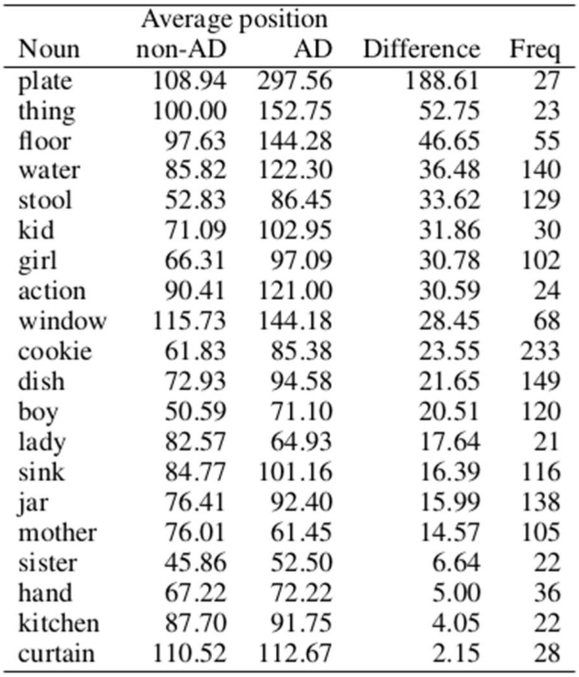
4.1. Training Set Analysis

In order to gain insight into whether the above hypothesis of temporal contrast between AD and non-AD patients is plausible, we conducted a statistical analysis on the training set, focusing on nouns, due to their function of denoting objects that can be easily connected to specific events in the image. Using the Stanza library (Qi et al., 2020) for assigning part-of-speech tags and lemmatisation, we extracted lemmas of nouns that appear at least 20 times in the test set. A threshold of 20 was used to filter out words used by a small minority of patients, which do not necessarily describe the events depicted in the picture. Since we are only interested in the differences between the target groups in regards to temporal aspect of the patient's description of the image, we also removed nouns that appear only in transcripts belonging to a certain group. This way we obtain a list of 20 nouns presented in **Figure 1**, which correspond to the constituents of the picture description task.

We determine a transcript position for each appearance of each noun (e.g., if the noun appears as the first word in the corpus, the position is one) and calculate an average noun position for each class, that is, the average of all positions of a specific noun in each class. The nouns in **Figure 1** are sorted according to the difference between the average positions in each class.

One can see that the noun *plate*, for instance, has very different positions in descriptions produced by the distinct groups. It appears in sentences such as *"Two cups and a plate are on the counter there."* and *"The lady is wiping a plate while the sink overflows"*., which are sentences describing details most likely not noticed by all participants (Cummings, 2019). Another noun with different positions is *thing*, which appears in sentences such as *"And the whole thing is going to collapse."*, describing more than just one specific element or an action concerning several constituents in the picture. *Floor*, the noun with the third biggest difference between the average positions in each class on the other hand appears in sentences such as *"There's water on the floor."* and *"And the stool is going to knock him on the floor."*, and is related to more central parts of the action seen in the picture. While both AD patients and non-AD control group use these nouns to describe the picture and the actions related to these nouns, they appear to focus on them at different times in their descriptions.





	Average position			
Noun	non-AD	AD	Difference	Freq
plate	108.94	297.56	188.61	27
thing	100.00	152.75	52.75	23
floor	97.63	144.28	46.65	55
water	85.82	122.30	36.48	140
stool	52.83	86.45	33.62	129
kid	71.09	102.95	31.86	30
girl	66.31	97.09	30.78	102
action	90.41	121.00	30.59	24
window	115.73	144.18	28.45	68
cookie	61.83	85.38	23.55	233
dish	72.93	94.58	21.65	149
boy	50.59	71.10	20.51	120
lady	82.57	64.93	17.64	21
sink	84.77	101.16	16.39	116
jar	76.41	92.40	15.99	138
mother	76.01	61.45	14.57	105
sister	45.86	52.50	6.64	22
hand	67.22	72.22	5.00	36
kitchen	87.70	91.75	4.05	22
curtain	110.52	112.67	2.15	28

FIGURE 1 | Average position of nouns that appear at least 20 times in the training set. AD and non-AD stand for average position in the speech transcripts of patients with AD and control group patients, respectively. Difference denotes the absolute difference between these averages, and Freq denotes the frequency of the noun in the corpus. The nouns are sorted according to the difference column.

The nouns at the end of the list are also interesting, since they denote situations in the picture described synchronously by both AD and non-AD patients. Noun *hand* is mostly related to a situation of the boy grabbing a cookie (e.g., *"He's grabbing a cookie in his hand."*) or to a situation of the girl reaching for a cookie (e.g., *"And the girl's trying to help and she's reaching her hand up."*). The noun *kitchen* appears in sentences such as *"Uh there's a set of kitchen cabinets."*, mostly describing static elements in the image. Similarly can be said for the noun *curtain*, which mostly appears in sentences describing static elements (e.g., *Curtains at the window.*) but can nevertheless also appear in sentences describing some rather detailed observations (*e.g., "Curtains are blowing I think."*).

4.2. Modelling Temporal Differences With Temporal Bag-of-Words

While the statistical analysis above offers some evidence of temporal differences in transcripts of AD and non-AD patients, a question remains as to whether classification models can detect these subtle differences. While assuming that they can is in our opinion a reasonable hypothesis, there is at least one reason to doubt this hypothesis. The presence of stronger features (i.e., semantic features, such as unigrams appearing only in one class) might cause the classifier to ascribe low importance to less subtle temporal aspects. Since in this section our focus is on ascertaining whether modelling of temporal aspects of the transcripts is possible rather than obtaining optimal accuracy (which is addressed in section 5.1), we can easily avoid this problem by restricting the classifier's model to contain only temporal features.

Therefore, we employed a simple bag-of-words model (Baeza-Yates and Ribeiro-Neto, 1999) to confirm or reject the hypothesis that the temporal differences between the non-AD and AD groups observed in the training set (see section 4.1) are relevant to the classification model. To track temporal order each transcript is divided into three sequential chunks of the same word length[2]. Words in each transcript belonging to the first chunk are given a suffix of _1, words belonging to the second chunk are given a suffix of _2, and words belonging to the third chunk are given a suffix of _3. This way the same words appearing in different sections of the transcript are distinguished by the bag-of-words model and we therefore obtain three features for each word, since in this bag-of-words model the same word with a different suffix is treated as a different word. Thus, we build a classifier that, rather than simply focusing on semantic differences (i.e., how many times a specific word appears in a specific transcript belonging to a specific group) also focuses on temporal differences (i.e., whether a specific word appears in a specific temporal chunk of a transcript belonging to a specific group). As we limited the word features in this model to the nouns appearing in **Figure 1**, the classifier is learnt to predict AD only on the basis of 60 features (i.e., 20 words from a list, each of

[2] The number of chunks was determined by finding the largest possible number of chunks where each set of chunks containing words from distinct positions in the text would contain at least one instance of each noun presented in **Figure 1**.

them with three distinct suffixes according to the position in the text) derived from 20 nouns, which appear in transcripts of both AD and non-AD patients.

We used the same classification approach as in our classification experiments described in detail in section 5.1, that is, we trained and tested 50 random forest classifiers (Breiman, 2001) with 50 trees of maximum depth 5 by employing leave-one-out cross validation (LOOCV) on the training set, each time using a different random seed. The predictions of these models on the training set were then used for majority voting in order to derive final predictions[3].

We measured the performance of the model by calculating accuracy according to the following equation:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{1}$$

where TP stands for true positive examples (i.e., examples that the classifier correctly assigned to the AD class), TN stands for true negative examples (i.e., examples that the classifier correctly assigned to the non-AD class), FP stands for false positive examples (i.e., examples that the classifier incorrectly assigned to the AD class) and FN for false negative examples, which the classifier incorrectly assigned to the non-AD class.

The final majority voting accuracy for LOOCV is 77.78%, which indicates that the model is capable of successfully leveraging temporal differences. The Scikit-learn library (Pedregosa et al., 2011) implementation of the algorithm used in this experiment allows to extract the importance of features based on a measure of "impurity." That is, when training a single decision tree, we can compute how much each feature contributes to decreasing the weighted impurity, in our case measured with Gini impurity (Breiman, 2001). In the case of random forests, we measured the averaged decrease in impurity over trees to derive a feature importance score for each feature. To increase reliability we averaged these scores for each feature across the ensemble of 50 random forest classifiers in order to obtain the final scores for each word.

The scores for the nouns analysed in section 4.1 are presented in **Table 1**. The hypothesis is that nouns exhibiting the most temporal dissimilarities between the AD and non-AD classes identified in section 4.1 will also be used by the classifier to distinguish between the classes, resulting in larger feature importance scores. In this case, the sum of all three scores for each noun would give indication that the specific word appears in different sections of the transcript depending on the class to which the transcript belongs.

By measuring the Pearson correlation between the sums of scores (see column labelled "Sum" in **Table 1**) and differences in average position (column labelled "Difference" in **Figure 1**), we however obtain a weak non-significant negative correlation of −0.15 with a p-value of 0.53, indicating a possibility that the classifier considers more fine-grained temporal information, which is not visible by just averaging words' positions in the text. For example, the noun *window*, which was identified by the classifier as the most important feature out of all nouns in the list, does show a considerate difference in average position between AD and non-AD classes, but nevertheless still appears somewhere in the middle of the list in **Figure 1**. The same is true of the noun *sink*, which was identified as the second most important feature. Slightly more consistency between rankings can be observed at the bottom of both lists, for example when observing the ranking for nouns *kitchen* and *sister*.

TABLE 1 | Feature importance of nouns in a random forest classifier according to its position in 1st, 2nd, or 3rd chunk of each transcript.

Noun	1st chunk	2nd chunk	3rd chunk	Sum
Window	0.09904	0.02905	0.01041	0.13849
Sink	0.06526	0.03472	0.01101	0.11099
Stool	0.06090	0.02709	0.01988	0.10787
Action	0.07408	0.00796	0.00591	0.08795
Curtain	0.03686	0.02560	0.01131	0.07377
Mother	0.02548	0.01984	0.01852	0.06384
Dish	0.02689	0.01874	0.00951	0.05514
Cookie	0.03305	0.01190	0.00929	0.05424
Water	0.03082	0.01380	0.00704	0.05167
Hand	0.02241	0.01573	0.00780	0.04594
Girl	0.01303	0.01129	0.00828	0.03260
Boy	0.01023	0.00914	0.00903	0.02840
Jar	0.01080	0.00957	0.00724	0.02762
Plate	0.01398	0.00489	0.00475	0.02362
Floor	0.00970	0.00700	0.00651	0.02322
Kid	0.00787	0.00773	0.00566	0.02126
Thing	0.00657	0.00624	0.00424	0.01705
Sister	0.00870	0.00484	0.00112	0.01465
Lady	0.00482	0.00364	0.00264	0.01110
Kitchen	0.00578	0.00263	0.00217	0.01057

Sum is the sum of all three scores.

5. AD DETECTION

The results of the temporal analysis in section 4 suggest that temporal differences in the descriptions can be detected in the transcripts and can also be successfully leveraged for detection of dementia by ML. Although it is doubtful that a classifier relying solely on temporal features would be able to achieve good performance, these features might improve AD detection when combined with other features. For this reason, in this section we explore a less specialised approach toward AD detection, which attempts to incorporate as many modalities and aspects of these modalities as possible. First, instead of focusing only on the textual information, we also incorporate several features extracted from audio modality, which are naturally time-based. As with audio, many aspects of the text

[3] Note that in this experiment we did not use term frequency-inverse document frequency (TF-IDF) weighting (Baeza-Yates and Ribeiro-Neto, 1999), as we did in the experiments in section 5.1 since we simply wanted the classifier to focus on binary differences between features (i.e., whether a specific temporal unigram appears in a transcript of a specific class or not).

modality are incorporated, including temporal, structural and semantic aspects.

5.1. Methodology

The main methodological steps of the proposed approach are described below, namely preprocessing, feature engineering and classification.

5.1.1. Preprocessing

For audio preprocessing, speech segmentation was performed on the audio files that met the above described selection criteria. The study only focuses on the participants' speech; therefore, the investigators' speech was excluded from further processing. We extracted the participants' speech utterances using the timestamps obtained through DementiaBank.

The manual transcripts in CHAT format (MacWhinney, 2019) were first converted into word and token sequences which represent what was actually produced in speech. For instance, the annotations 'w [x n]', which indicate that the word 'w' was repeated n times were replaced by n repetitions of w, punctuation marks and various comments annotated between'[]' were removed. Also removed were symbols such as (.), (..), (...),<, <, / and xxx, as well as all punctuation.

Next, the processed transcripts were force-aligned with the speech recordings using the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008), which labels the pauses between words with 'sp' and produces time stamps for each word and for each pause. The word time stamps allowed us to split audio recordings at the level of words/pauses and conduct acoustic feature extraction for each word. The volume of each word was normalised to the range $[-1:+1]$ dBFS. Volume normalisation helps in smoothing over different recording conditions, particularly variations in microphone placement in relation to the participant.

5.1.2. Feature Engineering

The main steps of the feature engineering procedure are presented in **Figure 2** and described below. The entire procedure can be divided into four main phases, generation and concatenation of audio and textual feature vectors, generation of six ADR features and selection of five distinct feature configurations.

The audio feature extraction was performed using the openSMILE v2.1 toolkit, which is an open-source software suite for automatic extraction of features from speech, widely used for emotion and affect recognition in speech (Eyben et al., 2010). In this research we opted to employ only the *eGeMAPS* (Eyben et al., 2016) feature set, which exhibited good performance in previous research (Haider et al., 2020). The *eGeMAPS* feature set corresponds to a basic set of acoustic features based on their potential to detect physiological changes in voice production, as well as theoretical significance and proven usefulness in previous studies (Eyben et al., 2016). It contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and slope V0 features, as well as their most common statistical functionals, for a total of 88 features per speech segment. Pearson's correlation test was performed to remove acoustic features that were significantly correlated with duration ($|R| > 0.2$) to remove any bias toward the duration of words for machine learning. A total of 72 eGeMAPS features were therefore selected.

Following voice feature extraction we generated text features corresponding to the same words using GloVe embeddings (Pennington et al., 2014) of size 50 (for pauses, we generate a vector of 50 zeros). The audio and text features were normalised separately to the $[0, 1]$ interval and concatenated to derive vectors of 122 features (72 audio features and 50 text features) corresponding to an audio-textual embedding for each word or pause. These vectors were then used in the ADR procedure for aggregation of words/pauses on the speaker level (Haider et al., 2020).

Note that in our implementation of ADR, we only loosely followed the original ADR algorithm, introducing several modifications. The procedure consists of the following steps:

1. **Clustering of feature vectors**: All word level feature vectors were aggregated into clusters using k-means clustering[4]. This is in contrast with the original implementation (Haider et al., 2020), which employed self-organising maps (SOM) clustering (Kohonen, 1990) but in line with the work done by Martinc and Pollak (2020).
2. **Generation of the ADR features**: The ADR feature vector is composed of several features, namely **cluster counts, duration, audio-textual word/pause embeddings, audio-textual centroid embeddings, audio-textual embedding velocity and audio-textual centroid velocity**. Note that the last four features were not employed in the original ADR (Haider et al., 2020) and are meant to also model the semantic aspects of the text input besides the temporal and structural properties of text and audio. Since the original ADR only modelled audio recordings, these features have not been used before. The following is a brief description of each of these features:

   - **Cluster counts**: Number of feature vectors in each cluster for each participant's audio recording, that is, a histogram of the number of words/pauses present in each cluster.
   - **Duration**: A histogram representation of word/pause utterance duration for each participant's audio recording. As the number and duration of segments varies for each audio recording, we normalised the feature vector by dividing it by the total duration of segments present in each audio recording.
   - **Audio-textual word/pause embeddings**: The audio-textual embeddings obtained for each participant were aggregated into a sequence. Principal component analysis (PCA)[5] was conducted on the embedding sequence in order to reduce the dimensionality of each embedding to 1. Finally the sequence is truncated to the length of 128 if the sequence is too long, or padded with zeros if the sequence is too short.

[4] We use the Scikit library (Pedregosa et al., 2011) implementation of the algorithm: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[5] Using the Scikit library implementation of the algorithm: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

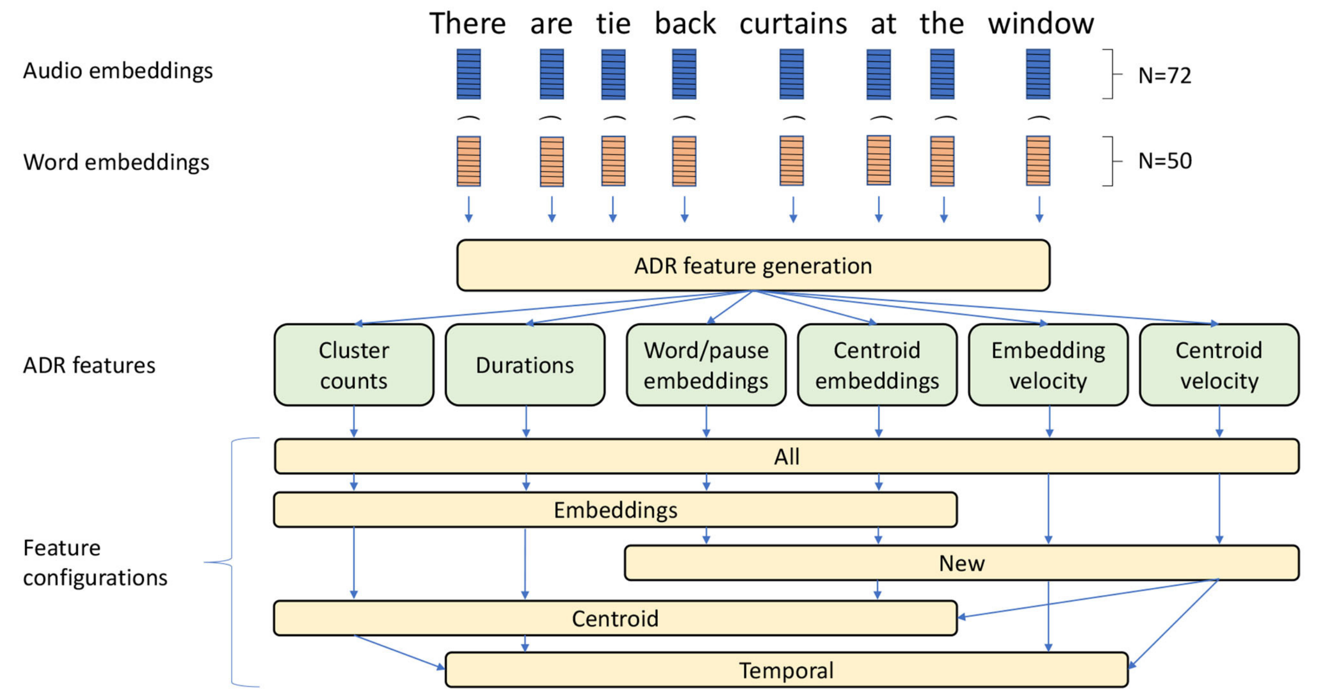


FIGURE 2 | Main feature engineering steps presented on the example preprocessed input sentence "There are tie back curtains at the window". Audio and word feature vectors (i.e., embeddings) are combined (Symbol "⌢" symbol denotes concatenations) and fed into an ADR feature generation procedure. The six resulting features are used in five distinct feature configurations.

At the end of this procedure, we obtained a vector of 128 features for each participant.

- **Audio-textual centroid embeddings**: Instead of employing PCA dimensionality reduction on audio-textual embeddings for each word, here we employed the procedure on the centroids of the clusters to which two consecutive word/pause utterances belong. At the end of this procedure, we obtained a vector of *k* features for each participant, where *k* is the number of clusters.
- **Audio-textual embedding velocity**: In order to model temporal aspects of speech and transcripts, we measured the change between consecutive audio-textual embeddings in the sequence. This is measured with cosine similarity between consecutive vectors t and e:

$$\cos(\mathbf{t},\mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n}\mathbf{t}_i\mathbf{e}_i}{\sqrt{\sum_{i=1}^{n}(\mathbf{t}_i)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{e}_i)^2}} \qquad (2)$$

The output of this procedure is a sequence of cosine distances between consecutive embeddings for each participant. The sequence was truncated (or padded with zeros) in order to obtain a vector of 128 features for each participant.

- **Audio-textual centroid velocity**: Similarly, change is measured with cosine similarity between cluster centroids to which two consecutive word/pause utterances belong. The resulting sequence of cosine similarities was again truncated (or padded with zeros) to the length of 128.

To establish the contribution of specific features and to gain a better sense of what type of information results in the best performance, we tested several feature configurations:

- **Temporal**: Includes only four ADR features that model only temporo-structural aspects of the audio and transcript data, namely cluster counts, duration, audio-textual embedding velocity and audio-textual centroid velocity.
- **Embedding**: Includes four ADR features that model structural and semantic aspects of the data, namely cluster counts, duration, audio-textual word/pause embeddings and audio-textual centroid embeddings.
- **Centroid**: Includes four ADR features that model structural, semantic and temporal aspects of the data, namely cluster counts, duration, audio-textual centroid embeddings and audio-textual centroid velocity.
- **New**: Includes only the four new ADR features which have not been used in the previous studies where ADR was employed (Haider et al., 2020; Martinc and Pollak, 2020), namely audio-textual centroid embeddings, audio-textual centroid velocity, audio-textual word/pause embeddings and audio-textual centroid embeddings.
- **All**: Includes all 6 ADR features described in section 5.1.2.

In addition, we investigated the impact of specific input modalities on the overall performance, or to be more specific, we employed three versions for each of the configurations above:

- **Audio**: Only audio input is used, consisting of a feature vector for each word/pause containing only 72 eGeMAPS features.
- **Text**: Only text input is used, that is, a feature vector for each word/pause containing only 50 GloVe embeddings features. Here, there are also no Duration features, which require audio recordings for its generation.
- **Text+audio**: Combination of text and audio features, consisting of a feature vector for each word/pause containing 122 eGeMAPS and GloVe embeddings features.

Finally, we investigated if performance could be improved by adding sub-word units consisting of four-character sequences (char4grams) into the model. Even with the additional ADR features for modelling semantic aspects of the text, the initial experiments still suggested that semantic modelling might be the biggest shortcoming of ADR. It is indeed possible that the compressed semantic information obtained from word embeddings by employing clustering, PCA or cosine similarity is not comprehensive enough, since it models semantics (or semantic change) only indirectly. To compensate for this and model semantics more directly, in some experiments we employ term frequency-inverse document frequency (TF-IDF) weighted word bound character char4gram features, which proved very successful at modelling semantics in the study by Martinc and Pollak (2020). Character n-grams are created only from text inside word boundaries and n-grams at the edges of words are padded with space[6].

5.1.3. Classification

To determine the best classifier for the task at hand and the best number of clusters (k), we first conducted a preliminary grid search across several classifiers and $k \in 10, 20, ..., 80$ values, in which we employed 5 classifiers from the Scikit library (Pedregosa et al., 2011), namely Xgboost (Chen and Guestrin, 2016) (with 50 gradient boosted trees with max depth of 10), random forest (with 50 trees of max depth of 5), SVM (with linear kernel and a box constraint configurations of 10), logistic regression (LogR, with a regularisation configuration of 10) and a linear discriminant analysis classifier. Only the *All* feature configuration was used during this preliminary experiment. Grid search was conducted on the training set, using LOOCV. Each classifier and k-value combination was run in the grid search five times, with five different random seeds for each classifier, in order to obtain more reliable results and to compensate for the observed variance in accuracy across different runs. The average accuracy across these five runs was used as a performance score for each combination of the classifier and k-value. Based on this score, the combination of k-means clustering with $k = 30$ and a random forest classifier was chosen for use in further experiments.

The large variance in accuracy (Equation 1) observed in these preliminary experiments is consistent with the observations of Yuan et al. (2020), where large variance in performance in the cross-validation setting was observed when employing BERT and ERNIE (Zhang et al., 2019) models. To solve this problem, they proposed a majority voting setting, in which the label assigned to an instance of the test set is the label assigned by the majority of the 50 models trained during cross-validation. We followed the same procedure and trained 50 models employing the same classifier and feature configuration on the training set, each time using a different random seed. These models were then used for majority voting on the test set to derive final predictions. The same procedure was employed to obtain comparable performance scores on the training set in LOOCV.

5.1.4. Baseline BERT Implementation

In order to conduct the temporal experiments reported in section 6.1 and obtain a strong baseline, we also leverage the BERT model (Devlin et al., 2019). The preprocessing employed here was as described above, treating pauses as a form of punctuation, following Yuan et al. (2020). The transcripts were then force-aligned with the speech signal, labelling pauses between words with “sp’, excluding pauses under 50 ms, and encoding short pauses (0.05–0.5 s) as ‘,’, medium pauses (0.5–2 s) as ‘.’, and long pauses (over 2 s) as ’...’.

In contrast to Yuan et al. (2020), we fed the processed transcripts to the pretrained ’bert-base-uncased’ language model with an additional linear sequence classification layer rather than the ’bert-large-uncased’ model. This was done so as to reduce the amount of computational resources required. We did not employ the ERNIE (Zhang et al., 2019) language model, since the publicly available implementation of the model[7] does not return the attention matrices required for the temporal analysis (see section 6.1). For fine-tuning, we employ the same hyperparameters as in the study by Yuan et al. (2020): learning rate = 2e-5, batch size = 4, epochs = 8, and maximum input length of 256. We set the standard BERT tokeniser not to split ’...’.

Finally, we once again employed majority voting both in the LOOCV setting and on the test set. Due to limited computational resources, we only conducted the LOOCV procedure five times, with five different seeds, therefore obtaining five predictions for each example in the training set. The majority vote of these five predictions is used as a final prediction. On the other hand, for the test set setting, we randomly choose 50 models out of 540 models generated during LOOCV and conduct majority voting on the predictions of these models to obtain the final predictions.

6. RESULTS

The results for the best feature combinations and input modalities are presented in **Table 2**. See **Supplementary Material** for a full table of results for all feature combinations (**Supplementary Table 2**), and for confusion matrices of the top 3 results and their late fusion (**Supplementary Figure 2**). For all results, we use k-means clustering with $k = 30$ and the random forest classifier, which yielded the best results in the preliminary grid search (see section 5.1.3).

Without late/decision fusion of the best three methods, the best result on the test set was achieved when *Temporal*

[6] For example, for the sentence *It is sunny today*, the following set of char4grams would be generated: {“It,” “is,” “sun,” “sunn,” “unny,” “nny,” “tod,” “toda,” “oday,” and “day.”}.

[7] https://github.com/PaddlePaddle/ERNIE

TABLE 2 | Results of the three best feature configurations in the LOOCV setting and on the test set in terms of accuracy.

Feature configurations	Input modality	LOOCV accuracy	Test set accuracy
Temporal + char4grams	audio + text	0.8611	**0.9167**
New + char4grams	audio + text	**0.8889**	0.8750
char4grams	text	0.8611	0.8958
top three late fusion	/	**0.8796**	**0.9375**
BERT—reimplementation of Yuan et al. (2020)	/	0.8426	0.8333
ERNIE best related work (Yuan et al., 2020)	/	/	0.8958

The feature configurations column indicates which feature configuration has been used and whether char4grams have been added, and column Input modality shows the modality on which ADR features have been generated. The best individual methods' results in LOOCV and on the test set, as well as the late fusion of all three methods, are shown in bold. The row labelled top three late fusion presents the results of employing late/decision fusion (i.e., the use of majority voting) over the three best approaches.

features generated on audio and text input were combined with *char4grams* (accuracy of 91.67%), and the best result in the cross-validation was achieved when *New* features generated on text and audio input were combined with *char4grams* (accuracy of 88.89%). *Char4grams* features by themselves also work very well, achieving an accuracy of 89.58% on the test set and accuracy of 86.11% in LOOCV. This indicates that semantic features and pause information contribute the most in terms of performance. Nevertheless, the results also indicate that we can improve the overall performance by including the temporal and structural aspects of audio and text.

Our reimplementation of BERT is noncompetitive in relation to the best approaches, reaching accuracy of 83.33% on the test set, which is in line with the results obtained by Yuan et al. (2020) who report accuracy of 85.4%. They however employ a larger BERT model with 24 layers and 16 attention heads for each layer.

The observations from the error analysis (see **Supplementary Material**) suggest that employing late fusion can be beneficial. In our experiments it improved the best achieved test set accuracy of 91.67% by about 2.3% (to 93.75%) despite a slight decrease in accuracy in the LOOCV setting (from 88.89% to 87.96%). Another beneficial improvement is due to the use of majority voting, which reduces the variability of the test set predictions of single classifiers, shown in **Figure 3**. **Figure 3** shows results of the accuracy distribution of 50 classifiers (employing temporal features and char4grams) used in the majority voting, when employed on the test set. It should be noted that the accuracy of 91.67% obtained by majority voting was obtained by <15% of classifiers in the ensemble, for the *temporal text+audio+char4grams* configuration. The other 85% of classifiers in the ensemble reach accuracy between 75 and 89%. **Figure 3** also shows that the spread is largest when only the audio modality is used, ranging from about 48% to almost 70%.

The approach presented in this paper outperformed all previous approaches to AD detection performed on this and similar spontaneous speech datasets, as shown in **Table 3**. All accuracy figures for text correspond to accuracy on manual transcripts. Of the studies shown in **Table 3**, only Mirheidari et al. (2018) report results for embeddings derived from ASR transcription (62.5% accuracy), in contrast to the 75.6% they obtained from manual transcription. As noted, comparisons of studies done on different subsets and training/test splits of the Pitt corpus are problematic. The best previous result on the same dataset (ADReSS) used in our study was achieved by Yuan et al. (2020), who reported 89.58% test-set accuracy obtained with an ensemble of ERNIE models. Our late fusion method yielded an improvement of about 4.7% over the best reported result on the ADReSS dataset, and an improvement of 25% over the ADReSS challenge baseline (Luz et al., 2020).

6.1. Dissecting the BERT Attention Space

Another way to gain insight into how temporal information can be leveraged for AD detection, is through the use of neural networks, which model temporal and structural dependencies by default. The baseline BERT implementation described in section 5.1.4 is based on the transformer architecture, which employs the attention mechanism. The attention mechanism can be analysed and visualised, offering insights into the inner workings of the system. BERT's attention mechanism consists of 12 attention heads (Vaswani et al., 2017)—square matrices linking pairs of tokens within a given text. We explored how this (activated) weight space can be further inspected to establish to what extend BERT models temporal information.

While square attention matrices show the importance of the correlations between all tokens in the transcript, we focused only on the diagonals of the matrices, which indicate how much attention the model pays to a specific word in relation to itself, giving a measure of how important a specific word is for the classification of a specific description as belonging to either the AD or the non-AD class.

As explained in section 5.1.4, the BERT model was fine-tuned through LOOCV on the training set, and the fine-tuning procedure resulted in 50 BERT models, which were used for prediction on the test set. We extracted diagonal attention scores for 12 attention heads for each of the 20 nouns presented in **Figure 1** appearing in different positions in different transcripts in the test set and averaged the scores across all 50 models. If a specific noun appeared in the same position in two or more different transcripts, scores belonging to the same position in each head were averaged. Finally, we also averaged the 12 attention heads scores for each position for each word so as to derive a sequence of attention scores for each noun. **Figure 4** presents these sequences of attention scores for each of the 20 nouns appearing in different positions in the transcript. The height of each column indicates the attention given to a specific noun at position in the transcript, and the colour of the column labels the class of the transcript, blue denoting the non-AD class and red denoting the AD class.

Figure 4 shows that BERT generally tends to focus more attention to nouns appearing at the beginning of the transcript and less attention to nouns appearing at the end of the transcript. For example, for the noun *curtain*, attention scores are skewed

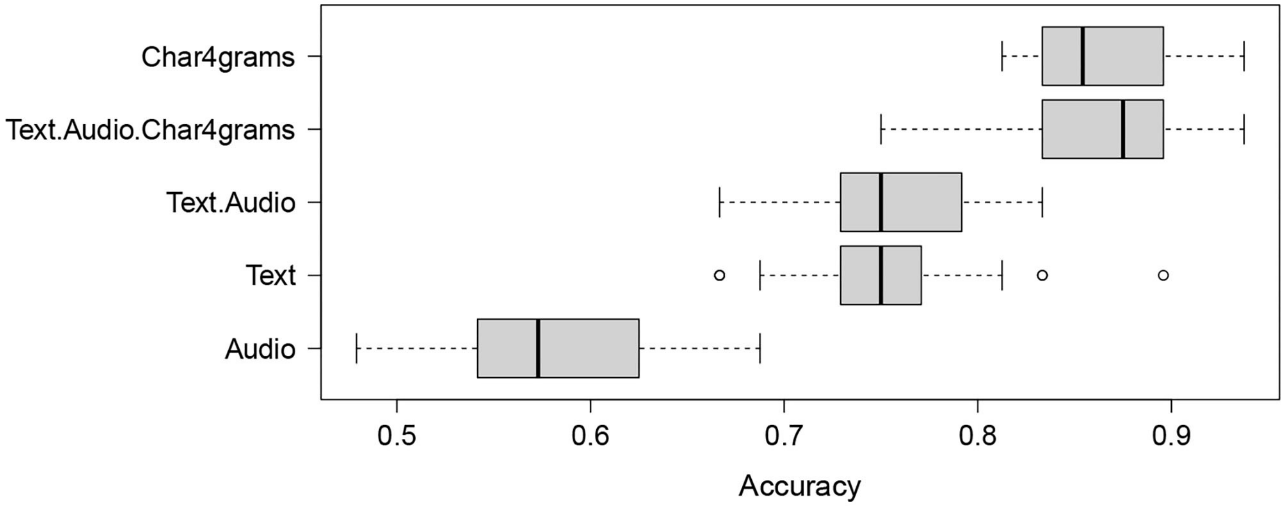


FIGURE 3 | Boxplot summarizing the accuracy distributions for 50 classifiers on the test set for the *Temporal* feature configuration (text, audio, and text+audio), char4grams and char4grams combined with text and audio (text+audio+char4grams).

TABLE 3 | Comparison with state-of-the-art studies conducted on subsets of the Pitt dataset.

Study	Accuracy	Modality
Haider et al. (2020)	78.7%	Acoustic
Luz (2017)	68.0%	Acoustic
Fraser et al. (2016)	81.9%	Text/acoustic
Yancheva and Rudzicz (2016)	80.0%	Text/acoustic
Hernández-Domínguez et al. (2018)	68.0%	Text
Mirheidari et al. (2018)	75.6%	Text
Studies based on the ADReSS dataset		
ADReSS challenge baseline	62.5%	Acoustic
ADReSS challenge baseline	75.00%	Text
Yuan et al. (2020) ERNIE	**89.58%**	Text
Yuan et al. (2020) BERT	85.40%	Text
Syed et al. (2020)	85.42%	Text
Balagopalan et al. (2020)	83.33%	Text
Sarawgi et al. (2020)	83.33%	Text/acoustic
Pompili et al. (2020)	81.25%	Text/acoustic
Koo et al. (2020)	81.25%	Text/acoustic
Cummins et al. (2020)	81.25%	Text/acoustic
Searle et al. (2020)	81.25%	Text/acoustic
Edwards et al. (2020)	79.17%	Text/acoustic
Rohanian et al. (2020)	79.17%	Text/acoustic
Martinc and Pollak (2020)	77.08%	Text
Pappagari et al. (2020)	75.00%	Text/acoustic
This study (best single model)	***91.67%***	Acoustic/text/temporal
This study (late fusion)	***93.75%***	Acoustic/text/temporal

The top three results are shown in bold. Results of this study are presented in Italics.

toward the first few appearances of the word, dropping drastically afterwards. This suggests that the appearance of the word curtain in the last part of the transcript is not important for classification. A similar pattern can be discerned for the nouns *sister* and *window*. It can also be observed that some words (e.g., *hand, floor, kitchen* and *plate*) are not given as much attention as others, regardless of the position at which they appear.

While the attention scores derived from BERT suggest that the position of the word in the AD classification task does matter, there is no clear correlation between the attention scores given by BERT and the difference in average position for specific words identified in section 4.1. This might indicate that identification of temporal aspects is somewhat more involved than hypothesised, depending not only on the words' position but also on the context in which it appears.

7. CONCLUSIONS

We presented a study of automatic detection of AD in spontaneous speech using state-of-the-art ML methods. We conducted a temporal analysis of the descriptions of the Cookie Theft scene of the Boston Diagnostic Aphasia Exam (Goodglass et al., 2001) in order to investigate putative temporal differences between descriptions produced by AD and non-AD patients, and to explore the modelling of these differences by ML. We then proposed a new AD detection approach, in which ADR is employed as a framework for multimodal feature extraction and fusion. Through this approach our model was able to surpass the best state-of-the-art results reported in the literature for the task of distinguishing between transcripts and audio recordings belonging to AD and non-AD participants in the ADReSS subset of the Pitt Corpus.

While our models were able to distinguish between AD and healthy controls with relatively high accuracy using spontaneous speech data, further validation on larger and more diverse datasets is warranted. As pointed out by de la Fuente Garcia et al. (2020), datasets suitable for AI studies of

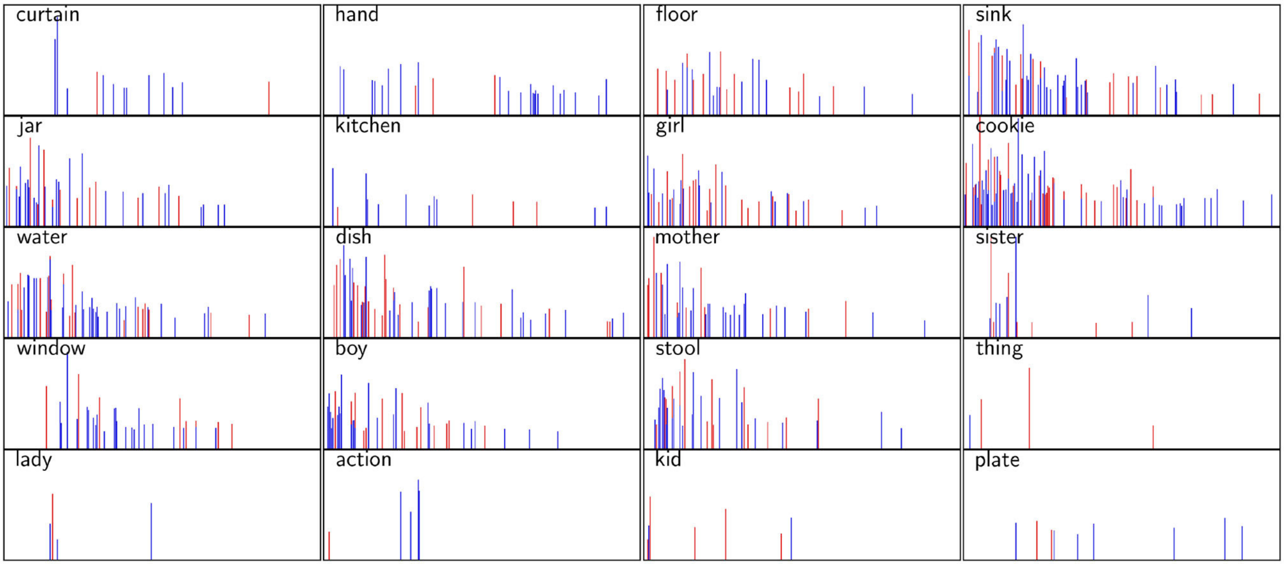


FIGURE 4 | Test set attention scores prescribed by the BERT model for 16 nouns presented in **Figure 1**. The height of each column indicates the attention given to a specific noun in a specific position in the transcript. A blue coloured column indicates that a specific noun appeared in the transcript belonging to the non-AD class, while a red coloured column indicates that a noun at this position appeared in a transcript belonging to the AD class. The positions (*x*-axis) range from 1 (i.e., first word in the transcript) to 256 (i.e., last word in the corpus).

the effects of neurodegeneration on spontaneously produced speech are relatively scarce at present. While this situation is changing, we hope our study will provide further impetus for research focused on elicitation and gathering of speech data from Alzheimer's cohort studies. An example of such studies is the PREVENT-ED spontaneous speech task, which has collected spontaneous dialogical speech from a group of healthy participants which includes participants genetically at-risk of AD, due to family history and apolipoprotein E (APOE) gene status (de la Fuente et al., 2019). Once the PREVENT-ED dataset has been fully collected, we aim to apply the methods presented in this article to investigate possible associations between speech features and the biomarkers available for the PREVENT cohort, including plasma and CSF Aβ42 amyloid, Tau and pTau, proinflammatory cytokines, acute-phase proteins, medial temporal-lobe atrophy and white matter lesion volume, as well as risk level (high, medium or low) and cognitive performance scores.

The results of the temporal bag-of-words model proved inconclusive in relation to the results of the analysis conducted in section 4.1. On the other hand, BERT results, while exhibiting sensitivity to temporal order, as words in different positions have different attention scores, are somewhat hard to interpret. These scores not only depend on temporal information but also indicate other differences between AD and non-AD patients related to semantic and grammatical contexts. We plan to address deficiencies of the temporal analysis and modelling in future work by investigating new temporal models and improving on our existing techniques for distillation of temporal information from the text.

Classification results indicate that accuracy gains can be achieved by adding temporal and structural information to semantic features. For example, the results show that the accuracy using only char4grams features (89.58%) can be improved to 91.67% when a combination of *temporal* audio textual features and char4grams features is employed, and up to 93.75% when late fusion of three best models is applied. These results compare favourably to the state-of-the-art. While these figures must be approached with caution given the relatively small size of the dataset, they provide motivation for further research into more challenging problems, such as earlier detection and prediction of AD progression, when suitable data become available in future.

Although the use of acoustic features on their own proved less successful than when combined with text, extraction of acoustic features can be fully automated, unlike textual features which if extracted through ASR would likely degrade classification accuracy. Therefore, while the multimodal approaches commonly employed in the recent ADReSS challenge (see **Table 3**) and extensively investigated in our study tend to benefit only marginally from the addition of acoustic information, processing and use of acoustic features is likely to remain an important topic of research in AD modelling, as will temporal aspects of spontaneous speech production.

As regards the use of transformer based embeddings, we believe they remain promising, despite their somewhat underwhelming contribution to classification performance in this study. Among other things, along with acoustic features, transformer based embeddings may play a role in the creation of language independent models for AD detection. Currently, multilingual BERT is being used in a variety of tasks allowing for classification to be performed on a language other than the language on which the model was trained ("zero-shot" transfer), and leveraging this possibility for cognitive decline detection would represent a valuable contribution to this field given that existing datasets are limited to only a few languages.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The data can be found at https://dementia.talkbank.org. The source code for the experiments in this paper is available at https://github.com/matejMartinc/alzheimer_diagnosis.

AUTHOR CONTRIBUTIONS

SL, SP, FH, and MM conceived and designed the experiments and analysis. FH and SL compiled and prepared the dataset. MM and FH performed the analysis. SL and SP drafted the initial manuscript. All authors contributed to the final version of the manuscript.

FUNDING

This work funded by the European Union's Horizon 2020 research and innovation program, under the grant agreement No. 769661, toward the SAAM project. We also acknowledge the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103).

ACKNOWLEDGMENTS

We thank Brian MacWhinney and Davida Fromm for creating and facilitating our access to DementiaBank.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2021.642647/full#supplementary-material

REFERENCES

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* 136, 3727–3737. doi: 10.1093/brain/awt269

Baeza-Yates, R., and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Harlow: Addison-Wesley-Longman.

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). "To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection," in *proceedings Interspeech 2020* (Shanghai), 2167–2171.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease. *Arch. Neurol.* 51:585. doi: 10.1001/archneur.1994.00540180063015

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 785–794.

Corey Bloom, J., and Fleisher, A. (2000). The natural history of Alzheimer's disease. *Dementia* 34, 405–415. doi: 10.1201/b13239-64

Cronin-Golomb, A. (2011). *Visuospatial Function in Alzheimer's Disease and Related Disorders*, Chapter 15, Wiley Online Library, 457–482.

Cummings, L. (2019). Describing the Cookie Theft picture: sources of breakdown in Alzheimer's dementia. *Pragmat. Soc.* 10, 153–176. doi: 10.1075/ps.17011.cum

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., et al. (2020). "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," in *Proceedings Interspeech 2020* (Shanghai), 2182–2186.

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J. Alzheimers Dis.* 78, 1547–1574. doi: 10.3233/JAD-200888

de la Fuente, S., Ritchie, C., and Luz, S. (2019). Protocol for a conversation-based analysis study: PREVENT-ED investigates dialogue features that may help predict dementia onset in later life. *BMJ Open* 9, 1–10. doi: 10.1136/bmjopen-2018-026254

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1* (Minneapolis, MN: Association for Computational Linguistics), 4171–4186. doi: 10.18653/v1/N19-1423

Dukart, J., Schroeter, M. L., Mueller, K., Alzheimer's Disease Neuroimaging Initiative, et al. (2011). Age correction in dementia–matching to a healthy brain. *PLoS ONE* 6:e22193. doi: 10.1371/journal.pone.0022193

Duong, A., Giroux, F., Tardif, A., and Ska, B. (2005). The heterogeneity of picture-supported narratives in Alzheimer's disease. *Brain Lang.* 93, 173–184. doi: 10.1016/j.bandl.2004.10.007

Edwards, E., Dognin, C., Bollepalli, B., and Singh, M. (2020). "Multiscale system for Alzheimer's dementia recognition through spontaneous speech," in *Proceedings Interspeech 2020* (Shanghai), 2197–2201.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2016). The Geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 190–202. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM-MM*, ACM, 1459–1462.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49, 407–422. doi: 10.3233/JAD-150520

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *BDAE-3: Boston Diagnostic Aphasia Examination–Third Edition*. Philadelphia, PA: Lippincott Williams & Wilkins.

Guo, Z., Ling, Z., and Li, Y. (2019). Detecting Alzheimer's disease from continuous speech using language models. *J. Alzheimers Dis.* 70, 1163–1174. doi: 10.3233/JAD-190452

Haider, F., de la Fuente, S., and Luz, S. (2020). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE J. Select. Top. Signal Proc.* 14, 272–281. doi: 10.1109/JSTSP.2019.2955022

Haulcy, R., and Glass, J. (2021). Classifying Alzheimer's disease using audio and text-based representations of speech. *Front. Psychol.* 11:3833. doi: 10.3389/fpsyg.2020.624137

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., and Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimers Dementia.* 10, 260–268. doi: 10.1016/j.dadm.2018.02.004

Hodges, J. R., and Patterson, K. (1995). Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. *Neuropsychologia* 33, 441–459. doi: 10.1016/0028-3932(94)00127-B

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., et al. (2014). "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Baltimore, MD: Association for Computational Linguistics), 27–37. doi: 10.3115/v1/W14-3204

Jonell, P., Moëll, B., Håkansson, K., Henter, G. E., Kuchurenko, T., Mikheeva, O., et al. (2021). Multimodal capture of patient behaviour for improved detection of early dementia: clinical feasibility and preliminary results. *Front. Comput. Sci.* 3:10. doi: 10.3389/fcomp.2021.642633

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480. doi: 10.1109/5.58325

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dementia* 1, 112–124. doi: 10.1016/j.dadm.2014.11.012

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). "Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition," in *Proceedings Interspeech 2020* (Shanghai), 2217–2221. doi: 10.21437/Interspeech.2020-3153

Lopez-de Ipiña, K., Alonso, J., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., et al. (2015). On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cogn. Comput.* 7, 44–55. doi: 10.1007/s12559-013-9229-9

Lopez-de Ipiña, K., Faundez-Zanuy, M., Solé-Casals, J., Zelarin, F., and Calvo, P. (2016). "Multi-class versus one-class classifier in spontaneous speech analysis oriented to Alzheimer disease diagnosis," in *Recent Advances in Nonlinear Speech Processing*, eds A. Esposito, M. Faundez-Zanuy, A. M. Esposito, G. Cordasco, T. D. J. Solé-Casals and F. C. Morabito, Vol. 48 (Springer International Publishing), 63–72.

Luz, S. (2009). "Locating case discussion segments in recorded medical team meetings," in *Proceedings of the ACM Multimedia Workshop on Searching Spontaneous Conversational Speech (SSCS'09)*, Beijing: ACM Press, 21–30.

Luz, S. (2017). "Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)* (Thessaloniki: IEEE), 45–46.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," in *Proceedings of Interspeech 2020* (Shanghai), 2172–2176.

Luz, S., la Fuente, S. D., and Albert, P. (2018). "A method for analysis of patient speech in dialogue for dementia detection," in *Proceedings of LREC'18*, ed D. Kokkinakis (Paris: ELRA), 35–42.

MacWhinney, B. (2019). Understanding spoken language through talkbank. *Behav. Res. Methods* 51, 1919–1927. doi: 10.3758/s13428-018-1174-9

Mahajan, P., and Baths, V. (2021). Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. *Front. Aging Neurosci.* 13:20. doi: 10.3389/fnagi.2021.623607

Martinc, M., and Pollak, S. (2020). "Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia," in *Proceedings of Interspeech 2020*, (Shanghai), 2157–2161.

Meguro, K., Shimada, M., Someya, K., Horikawa, A., and Yamadori, A. (2001). Hemispatial visual-searching impairment correlated with decreased contralateral parietal blood flow in Alzheimer disease. *Cogn. Behav. Neurol.* 14, 213-218.

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019a). Dementia detection using automatic analysis of conversations. *Comput. Speech Lang.* 53, 65–79. doi: 10.1016/j.csl.2018.07.006

Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2018). "Detecting signs of dementia using word vector representations," in *Interspeech*, 1893–1897.

Mirheidari, B., Pan, Y., Walker, T., Reuber, M., Venneri, A., Blackburn, D., et al. (2019b). Detecting Alzheimer's disease by estimating attention and elicistation path through the alignment of spoken picture descriptions with the picture prompt. *arXiv [Preprint]. arXiv:1910.00515.*

Orimaye, S. O., Wong, J. S., Golden, K. J., Wong, C. P., and Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinform.* 18:34. doi: 10.1186/s12859-016-1456-0

Pappagari, R., Cho, J., Moro-Velàzquez, L., and Dehak, N. (2020). "Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity," in *Proceedings Interspeech 2020*, 2177–2181.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.org/papers/v12/pedregosa11a.html

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha), 1532–1543.

Petti, U., Baker, S., and Korhonen, A. (2020). A systematic literature review of automatic alzheimer's disease detection from speech and language. *J. Am. Med. Inform. Assoc.* 27, 1784–1797. doi: 10.1093/jamia/ocaa174

Pompili, A., Rolland, T., and Abad, A. (2020). "The INESC-ID multi-modal system for the ADReSS 2020 challenge," in *Proceedings Interspeech 2020* (Shanghai), 2202–2206.

Pope, C., and Davis, B. H. (2011). Finding a balance: the carolinas conversation collection. *Corpus Linguist. Linguist. Theory* 7, 143–161. doi: 10.1515/cllt.2011.007

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: a python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*. doi: 10.18653/v1/2020.acl-demos.14

Reilly, J., Rodriguez, A. D., Lamy, M., and Neils-Strunjas, J. (2010). Cognition, language, and clinical pathological features of non-Alzheimer's dementias: an overview. *J. Commun. Dis.* 43, 438–452. doi: 10.1016/j.jcomdis.2010.04.011

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans. Audio Speech Lang. Proc.* 19, 2081–2090. doi: 10.1109/TASL.2011.2112351

Rohanian, M., Hough, J., and Purver, M. (2020). "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," in *Proceedings Interspeech 2020* (Shanghai), 2187–2191.

Sarawgi, U., Zulfikar, W., Soliman, N., and Maes, P. (2020). "Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity," in *Proceedings Interspeech 2020* (Shanghai), 2212–2216.

Searle, T., Ibrahim, Z., and Dobson, R. (2020). "Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech," in *Proceedings Interspeech 2020* (Shanghai), 2192–2196.

Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). "Automated screening for Alzheimer's dementia through spontaneous speech," in *Proceedings Interspeech 2020* (Shanghai), 2222–2226.

Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., et al. (2017). Detecting dementia through interactive computer avatars. *IEEE J. Trans. Eng. Health Med.* 5:2200111. doi: 10.1109/JTEHM.2017.2752152

Toth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous Speech. *Curr. Alzheimer Res.* 15, 130–138. doi: 10.2174/1567205014666171121114930

Turner, R. S., Stubbs, T., Davies, D. A., and Albensi, B. C. (2020). Potential new approaches for diagnosis of alzheimer's disease and related dementias. *Front. Neurol.* 11:496. doi: 10.3389/fneur.2020.00496

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Vighetto, A. (2013). Towards an earlier diagnosis of Alzheimer's disease presenting with visuospatial disorders (posterior cortical atrophy). *Revue Neurol.* 169, 687–694. doi: 10.1016/j.neurol.2013.08.001

Weiner, J., Frankenberg, C., Schröder, J., and Schultz, T. (2019). "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (Sentosa: IEEE), 674–681.

Weiner, J., Herff, C., and Schultz, T. (2016). "Speech-based detection of Alzheimer's disease in conversational German," in *17th Annual Conference of the International Speech Communication Association* (ISCA) (San Francisco, CA), 1938–1942.

Yancheva, M., and Rudzicz, F. (2016). "Vector-space topic models for detecting Alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1, Long Papers* (Berlin), 2337–2346.

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease," in *Proceedings of Interspeech 2020*, (Shanghai), 2162–2166.

Yuan, J., and Liberman, M. (2008). Speaker identification on the scotus corpus. *J. Acoust. Soc. Am.* 123:3878. doi: 10.1121/1.2935783

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). "ERNIE: Enhanced Language Representation with Informative Entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), 1441–1451. doi: 10.18653/v1/P19-1139

Conflict of Interest: The 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.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 18 June 2021
doi: 10.3389/fcomp.2021.640669

Alzheimer's Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features

Shamila Nasreen[1,2]*, *Morteza Rohanian*[1], *Julian Hough*[1] *and Matthew Purver*[1,3]

[1]Cognitive Science Group, School of Electronic Engineering and Computer Science (EECS), Queen Mary University of London, London, United Kingdom, [2]Department of Software Engineering, Mirpur University of Science and Technology, Mirpur, Pakistan, [3]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Peter Garrard,
St George's, University of London,
United Kingdom
Daniel Blackburn,
Sheffield Teaching Hospitals NHS Foundation Trust, United Kingdom
Anna Pompili,
Other, Portugal

Correspondence:
Shamila Nasreen
shamila.nasreen@qmul.ac.uk

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *11 December 2020*
Accepted: *12 May 2021*
Published: *18 June 2021*

Citation:
Nasreen S, Rohanian M, Hough J and Purver M (2021) Alzheimer's Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features. Front. Comput. Sci. 3:640669. doi: 10.3389/fcomp.2021.640669

Alzheimer's disease (AD) is a progressive, neurodegenerative disorder mainly characterized by memory loss with deficits in other cognitive domains, including language, visuospatial abilities, and changes in behavior. Detecting diagnostic biomarkers that are noninvasive and cost-effective is of great value not only for clinical assessments and diagnostics but also for research purposes. Several previous studies have investigated AD diagnosis via the acoustic, lexical, syntactic, and semantic aspects of speech and language. Other studies include approaches from conversation analysis that look at more interactional aspects, showing that disfluencies such as fillers and repairs, and purely nonverbal features such as inter-speaker silence, can be key features of AD conversations. These kinds of features, if useful for diagnosis, may have many advantages: They are simple to extract and relatively language-, topic-, and task-independent. This study aims to quantify the role and contribution of these features of interaction structure in predicting whether a dialogue participant has AD. We used a subset of the Carolinas Conversation Collection dataset of patients with AD at moderate stage within the age range 60–89 and similar-aged non-AD patients with other health conditions. Our feature analysis comprised two sets: *disfluency* features, including indicators such as self-repairs and fillers, and *interactional* features, including overlaps, turn-taking behavior, and distributions of different types of silence both within patient speech and between patient and interviewer speech. Statistical analysis showed significant differences between AD and non-AD groups for several disfluency features (edit terms, verbatim repeats, and substitutions) and interactional features (lapses, gaps, attributable silences, turn switches per minute, standardized phonation time, and turn length). For the classification of AD patient conversations vs. non-AD patient conversations, we achieved 83% accuracy with disfluency features, 83% accuracy with interactional features, and an overall accuracy of 90% when combining both feature sets using support vector machine classifiers. The discriminative power of these features, perhaps combined with more conventional linguistic features, therefore shows potential for integration into noninvasive clinical assessments for AD at advanced stages.

Keywords: Alzheimer's disease, spontaneous speech, disfluency, interaction, natural language processing, mental health monitoring

INTRODUCTION

Alzheimer's disease (AD) is a chronic neurodegenerative disorder of the brain and the most prevalent form of dementia. According to the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA), the most common symptoms include an inability to function at work or to perform usual activities, reduced cognitive capabilities (including impaired reasoning and visuospatial abilities, impaired ability to acquire and remember new information, impaired language function), and changes in behavior. Language deficit primarily occurs through a decline in lexical semantic abilities with anomia and word comprehension, object naming, semantic paraphasias, and a decrease in vocabulary and verbal fluency throughout the entire span of the disease (Bayles and Boone, 1982; Forbes-McKay and Venneri, 2005). Effects are also seen at the pragmatic level, including problems with maintaining and alteration in discourse planning (Chapman et al., 2002). At the phonetic and phonological level, speech in patients with AD is principally characterized by a low speech rate and by frequent hesitations (Hoffmann et al., 2010); however, syntactic processing is relatively preserved at the early stages of the disease (Kavé and Levy, 2003; Forbes-McKay and Venneri, 2005).

There is no single universally accepted medical test for the diagnosis of AD; instead, physicians typically use a variety of methods with the help of specialists (including neurologists) to make a diagnosis. This includes a combination of taking feedback from family members and carers asking about changed patterns in behaviors and thinking, getting family history, and mental status examination. NINCDS established the criteria for AD diagnosis and requires that the presence of cognitive impairment needs to be confirmed by neuropsychological testing for a clinical diagnosis of possible or probable AD (McKhann et al., 1984). Neuropsychological testing should be performed when the routine history and bedside mental status examination cannot provide a confident diagnosis (McKhann et al., 2011). Suitable neuropsychological tests include the Mini-Mental Status Examination (Folstein et al., 1975), Mini-Cog (Rosen et al., 1984), Addenbrooke's Cognitive Examination–Revised (ACE-R) (Noone, 2015), Hopkins Verbal Learning Test (HVLT) (Brandt, 1991), and DemTect (Kalbe et al., 2004). Other routes include the use of blood tests and/or brain imaging (MRI) to check for high levels of beta-amyloid, an accumulation of protein fragments outside neurons, and one of the several brain changes associated with AD (Straiton, 2019).

Medical diagnoses based on the clinical interpretation of patients' history, complemented by brain scanning (MRI), are time-consuming, stressful, costly, and often cannot be offered to all patients complaining about functional memory. The other alternatives are extensive neurological screening tests that are used for the early diagnosis of AD and dementia. These tests require experts to interpret the results, strongly relying on brief cognitive tests, and are performed in medical clinics, with patients required to visit the clinics for diagnosis. There is a need for new, less invasive approaches that improve and speed up the process of early diagnosis, reduce distress to patients, and place less emphasis on extensive and expensive formal testing. Currently, researchers are therefore investigating the impact of neurodegenerative impairment on patients' speech and language, with the hope of deriving tests that are easier to administer and automate via natural language processing techniques (see, e.g., Fraser KC. et al., 2016).

Conversational dialogue is the primary means of human natural language use, so dialogue, and open domain dialogue in particular, might provide more generally applicable insights in studying the effects of AD on dialogue (Nasreen et al., 2019). Conversational analysis (CA) studies have traditionally looked in more detail at what characteristics of dialogue with dementia might be important (Jones et al., 2016; Elsey et al., 2015; Hamilton, 2005; Davis and Maclagan, 2010; Mirheidari et al., 2019; Perkins et al., 1998; Varela Suárez, 2018). Although some computational works explore the detection of dementia from speech and interaction (e.g. Luz et al., 2018; Broderick et al., 2018; Mirheidari et al., 2019), it is so far relatively limited, and there is little work on how dementia might affect interactional patterns in natural conversations (Addlesee et al., 2019).

AD is associated with many characteristic changes in language and speech not only with individual capabilities but also consequently in the interactive patterns observed in conversations. However, most language-based approaches so far use picture description or narrative tasks, or analyze individual speech, and thus miss conversational clues. This article examines the function of combining single-speaker disfluency features with interactional (dialogue) features to analyze the predictive power of these features in the diagnosis of AD. Extracts from the spontaneous speech of 15 AD and 15 non-AD patients from a conversational dataset, the Carolinas Conversation Collection (CCC), are analyzed to highlight the function of these interactional patterns, particularly pauses within a patient's utterances and during turn changes with a conversation partner in natural conversation. As will be described, we show the value of both disfluency and interactional information in conversation, combining them to achieve an overall accuracy of 90% in the recognition of AD from dialogue data.

PREVIOUS WORK

Much of the work to date in AD diagnosis has focused on properties of individual language, using various kinds of linguistic and acoustic features (Jarrold et al., 2014), or fluency, information content, and syntactic complexity (Fraser et al., 2016b; Fraser et al., a; de Lira et al., 2011). However, this is often studied within particular individual language tasks, usually within specific domains including picture description [the commonly used Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001)], story narration task [e.g. The Dog story (Le Boeuf, 1976)], and semi-structured interviews [e.g. Autobiographical Memory Interview (Kopelman et al., 1990)]. Approaches to analysis and diagnosis therefore usually focus on aspects of individual language such as lexical, grammatical, and semantic features. Kavé and Dassa (2018), for example, examined dementia via a picture description task in the Hebrew language, using ten linguistic features, and showed that the AD group produced a smaller percentage of content words, more pronouns relative to nouns and pronouns, a lower type-token ratio, and more frequent words as compared with cognitively intact participants. Orimaye et al. (2017) built an automated diagnosis model using low-level linguistic features including lexical, syntactic, and semantic features (NGrams) from verbal utterances of Probable AD and control participants. In another line of research, Ahmed et al. (2013) argued that speech production, syntactic complexity, lexical content, semantic content, idea efficiency, and idea density are important features of connected speech that are used to examine longitudinal profiles of impairment in AD.

Fluency has also been shown to be indicative of AD. Patients with AD have difficulty performing tasks that leverage semantic information, and exhibit problems with verbal fluency and identification of objects (Pasquier et al., 1995; López-de Ipiña et al., 2013). The semantics and pragmatics of their language appear affected throughout the entire span of the disease more than syntax (Bayles and Boone, 1982). Patients with AD talk more gradually with longer pauses and invest extra time seeking the right word, which contributes to disfluency of speech (López-de Ipi et al., 2013). Abel et al. (2009) modeled patient speech errors (naming and repetition disorders) to the problem of AD diagnosis. Rohanian et al. (2020) used a deep multi-modal fusion model to show the predictive power of disfluency features in the identification of AD.

Pausing behavior is often associated with a lack of fluency, and several studies have suggested various temporal forms of speech analysis to identify AD. During speech production, pauses are often considered a hallmark of a patient's lexical-semantic decline, one of the earliest symptoms of AD (Pistono et al., 2019b). Davis and Maclagan (2010) examined the silent pauses in a story retelling task with an older woman on two different occasions and found changes in pauses function signaling difficulty in word finding to difficulty in finding key component in the thread of a story. Forbes-McKay and Venneri (2005) compared the word-finding difficulties during the discourse in a picture description task among AD and healthy elderly subjects and stressed the fact that pauses, use of indefinite terms, and repetition are significantly more frequent in the AD group. According to Gayraud et al. (2011), AD patients produce more silence pauses than healthy controls but they found no significant difference in the duration of pauses. This study was performed on spontaneous speech data of an autobiographical task of AD and healthy persons and also identified that silent pauses occur more often outside syntactic boundaries and are followed by more frequent words. Singh et al. (2001) utilized different temporal measures including frequency of pauses, total pause time, mean duration of pause (MDP), standardized pause rate (SPR), standardized phonation time (SPT), and a few more to distinguish between AD and healthy control group by performing statistical analysis and discriminant analysis.

From a more linguistic perspective, silences in conversation have been analyzed in terms of distinct categories, with several terms coined to distinguish these, especially pauses at speaker changes or turn changes. Sacks et al. (1978) distinguished three kinds of silences in speech: pause (silence within the same speaker), gap (shorter silence at speaker change), and lapse (longer pause at speaker change). A normal gap duration is 200–1000 ms, as reported in the literature (Heldner and Edlund, 2010). Levinson (1983) employed a turn-taking system by integrating its forms and functions and categorized silence into three categories: within-turn silence (pause), inter-turn silence (gap or lapse), and turn silence (attributable silence). Researchers investigated turn silences within the framework of conversational analysis (CA) and Relevance Theory (RT) by taking into account the communicators' psychological factors, i.e. why they resort to silence rather than other means of communication to avoid giving a dis-preferred response (Wang, 2019). Applying these ideas to Alzheimer's discourse, Davis and Maclagan (2009) showed that both filled and silent pauses are keyed to functions within narration and within a conversation. They demonstrated that filled pauses (e.g. "uh" and "um") serve as placeholders and hesitation markers while silent pauses serve as a function for word finding, planning a word, and narrative level as well as an indicator of decreases in other interactional and narrative skills. They utilized the convention of Crystal and Davy (2016) to distinguish between micro-pause (less than a second), average pause (less than 2 s), and long pause (longer than 2 s) with elderly people (speech rate decreases with age).

CA's emphasis on conversation as a collaborative achievement demonstrates that examining interaction can provide more insight than separate analysis of the contributions of the two halves: each contribution to the conversation is built upon and responds to the partner's previous contribution. Perkins et al. (1998) explored turn-taking behavior, repairs, and topic management in conversations with dementia, and demonstrated that cognitive deficits may compromise the ability to secure the conversational floor or hold onto it and that failure to maintain topics often leads to topic changes by the conversational partner. Jones et al. (2016) presented a CA study of dyadic communication between clinicians and patients during initial specialist clinic visits, while Elsey et al. (2015) highlighted the role of carer, looking at triadic interactions among a clinician, a patient, and a carer. They established differential conversational profiles that distinguish between nonprogressive functional memory disorder (FMD) and progressive neurodegenerative disorder (ND), based on the interactional behavior of patients responding to neurologists' questions about their memory problems. Davis et al. (2014) examined how effective communication can be with the usage of strategies such as quilting, go ahead, and indirect questions between residents with dementia and their conversation partners, exploring various aspects including the impact of different types of questions, delayed responses, and the number of ideas in response using idea density.

Interactional features, therefore, promise one way to help alleviate the problems discussed in **Section 1**, by contributing to general, noninvasive methods of diagnosis that can be applied in natural everyday conversation, and some recent work has therefore investigated computational models using machine learning techniques. In a recent study, Mirheidari et al. (2019) performed an automated analysis for dementia detection with CA-inspired features, together with some language and acoustic features, achieving a classification accuracy of 90%. Luz et al. (2018) built a predictive model based on content-free features extracted from dialogue interactions from spontaneous speech in more natural settings using the CCC corpus of patient interview dialogues (Pope and Davis, 2011). They achieved promising results with an accuracy of 86% with only dialogue interaction-based features with less reliance on the content of task/dialogue. In a study building on the PREVENT Dementia project, de la Fuente Garcia et al. (2019) built a protocol for a conversation-based analysis study to investigate whether early behavioral signs of AD may be detected through dialogue interactions. Interactional patterns are considered among the current challenges to be addressed to make the spoken dialogue systems usable by older adults or frail patients (Addlesee et al., 2019). The purpose of this study is to investigate a new set of interactional features in AD conversations and evaluate their use in a computational model for AD classification.

DATASET AND FEATURES

Dataset and Participants

This study aims to investigate the behavior of AD patients based on the interaction patterns, including repairs and pauses within utterances and between turns, observed in a corpus of dialogue. This is a post hoc study based on an existing dataset, the CCC corpus, collected and distributed by the Medical University of South Carolina (MUSC) (Pope and Davis, 2011). The CCC corpus is a digital collection of semi-structured interviews including time-aligned transcripts with audio and video for some of the samples. These conversations are not based on a fixed task like picture description, but rather are based on the general discussion on daily routine, health, and different occasions like Christmas. AD subjects were aged 65 years and older with their AD at relatively moderate stages, while non-AD subjects include unimpaired persons with 12 chronic diseases of similar age. Each patient is interviewed by a different interviewer, either a linguistics student or a person from the community center involved. The demographic and clinical variables available include age range, gender, occupation prior to retirement, diseases diagnosed, and level of education (in years). Patients and interviewers are anonymized for security and privacy reasons. Access to the data was granted after ethical review by the both Queen Mary University of London (*via* QMERC 2019/04 dated April 25, 2019) and MUSC. As this dataset includes only elder patients, with diagnosed dementia of Alzheimer's type at moderate stage, it can only allow us to observe patterns associated with AD at a relatively advanced stage. This does not directly tell us whether these extend to early-stage diagnosis. However, it has the advantage of containing relatively free conversational interaction, compared to the more formulaic tasks and one-sided interaction available in corpora more commonly used in AD research, e.g. DementiaBank (Becker et al., 1994).

TABLE 1 | Demographic data for AD and non-AD patients, with dialogue duration in minutes.

	AD	Non-AD
	(*N* = 15)	(*N* = 15)
Age range	60–89	60–79
Years of education	9–16	8–16
Gender	M:4	M:4
–	F:11	F:11
Total duration of dialogues	152	179.7
Average dialogue duration	10.13	11.97

For this particular study, we use the transcript and audio recording from one dialogue conversation chosen randomly from each of a total of 30 patients: 15 AD diagnosed patients (4 male, 11 female) and 15 patients (4 male, 11 female) with other chronic diseases including diabetes, heart problems, arthritis, high cholesterol, cancer, leukemia but not AD; no patients were diagnosed as having breathing problems. These groups are selected to match the age range, to compare the different patterns of interaction, and to avoid bias. The demographic data of the participants are given in **Table 1**.

Disfluency Features

Detailed language use research helps us to find the indications of language impairment in AD and is a step toward the design of future clinical diagnostic tools. Disfluencies like self-repairs, pauses, and fillers are widespread in everyday speech (Schegloff et al., 1977). Disfluencies are usually seen as indicative of communication problems, caused by production or self-monitoring issues (Levelt, 1983). Individuals with AD are likely to deal with troubles in language and cognitive skills. Patients with AD speak more slowly and with longer breaks, and invest extra time seeking the right word, which in effect contributes to disfluency (López-de Ipi et al., 2013). The present research explores the disfluencies present in the speech of AD patients as they contribute to the severity of symptoms.

Self-repair disfluencies are typically assumed to have a reparandum–interregnum–repair structure, in their fullest form as speech repairs (Shriberg, 1994). A reparandum is a speech error subsequently fixed by the speaker; the corrected expression is a repair. An interregnum word is a filler or a reference expression between the words of repair and reparandum, often a halting step as the speaker produces the repair, giving the structure as in **(1)**

$$\textit{John}\ \underbrace{[\,\textit{likes}}_{\textit{reparandum}}\ +\ \underbrace{\{\,\textit{uh}\,\}}_{\textit{interregnum}}\ \underbrace{\textit{loves}\,]}_{\textit{repair}}\ \textit{Mary} \qquad (1)$$

In the absence of reparandum and repair, the disfluency reduces to an isolated ***edit term***. A marked, lexicalized edit term such as a filled pause ("uh" or "um") or more phrasal terms like "I mean" and "you know" can occur. Recognizing these elements and their structure is then the task of disfluency detection.

TABLE 2 | The proposed disfluency feature set.

Feature	Description
Patient features	
# *edit_terms*	Number of # *edit_terms* within *P* utterances normalized by the total # of words spoken by *P*
# *Rpt*	Number of verbatim repeats within *P* utterances normalized by the total # of words spoken by *P*
# *Sub*	Number of substitutions within *P* utterances normalized by the total # of words spoken by *P*
# *Del*	Number of deletes within *P* utterances normalized by the total # of words spoken by *P*
Interviewer features	
# *edit_terms*	Number of # *edit_terms* within *I* utterances normalized by the total # of words spoken by *I*
# *Rpt*	Number of verbatim repeats within *I* utterances normalized by the total # of words spoken by *I*
# *Sub*	Number of substitutions within *I* utterances normalized by the total # of words spoken by *I*
# *Del*	Number of deletes within *I* utterances normalized by the total # of words spoken by *I*

Here, each word is either tagged as a repair onset tag (marking the first word of the repair phase), edit term (*edit_terms*), or fluent word by the disfluency detector. To get the most information from different types of disfluency, we split repairs between the broad classes of verbatim repeats (***Rpt***), substitutions (***Sub***), and deletes (***Del***):

1) "So (he + he) brings the fresh flowers . . ."
Repeats
2) "(Someone said that + I heard someone out here say) it is getting quite cool outside, is it?"
Substitution
3) ". . .and I looked [at + (uh)] and answered her question. . ."
Deletes

We automatically annotated self-repairs using a deep-learning-driven model of incremental detection of disfluency developed by Rohanian and Hough (2020) and Hough and Schlangen (2017).[1] It consists of a deep learning sequence model, a long short-term memory (LSTM) network, which uses word embeddings of incoming words, part-of-speech annotations, and other features in a left-to-right, word-by-word manner to learn a sequence model of, and predict, disfluency tags according to the structure in **(1)** and any other edit term words. The model is trained on the disfluency detection training section of the Switchboard corpus (Godfrey et al., 1992), a sizable multispeaker corpus of conversational speech. Rohanian and Hough (2020) reported the automatic disfluency detector achieves an F1-score accuracy on detecting the first word of the repair phase at 0.743 and an F1-score accuracy of 0.922 on detecting all edit term words on the Switchboard disfluency detection test data. We considered its accuracy adequate for our purposes. Automatically deriving the types of interest from the tagger's output, we use four disfluency tags for patients (*P*) and four for interviewers *(I)* resulting in a total of eight disfluency features (details in **Table 2**).

Interactional Features

Annotation Protocol

We consider any silence of at least 0.5 s length for this particular study. To categorize the silences, we employed Levinson (1983)'s definitions: *pauses* (silences within a single speaker's turn), *gaps* and *lapses* (silences between speaker turns), and *attributable silences* (silences where speaker changes were expected but did not occur). We further categorized pauses into *short pause* (***SP***) and *long pause* (***LP***). An *SP* is a silence that occurs inside a single speaker turn, which we advised in the annotation protocol for average speech rates is greater than 0.5 s and less than 1.5 s; an *LP* is a longer pause within a single speaker turn, normally at least 1.5 s. We used guidelines for these thresholds rather than strict rules, because of different speech rates, and the judgment was left to annotators as to which category the pause fell into based on their perception. Both SPs and LPs may occur either at a *transition relevance place* (TRP) or not at a TRP, but no speaker change occurred. TRPs are junctures at which the turn could pass from one speaker to another.

TABLE 3 | Inter-annotator agreement: Cohen's kappa (κ) and observed agreement (A_o)

Feature name	Acronym	κ	A_o
Short pause	*SP*	0.55	0.83
Long pause	*LP*	0.46	0.79
Gap	*GA*	0.88	0.94
Lapse	*LA*	0.75	0.96
Attributable silence	*AS*	0.66	0.98
Overall	–	0.66	0.75

For inter-turn silences and attributable silences, we did not use explicit time thresholds—annotators used their judgment when listening to the silences in the context of the conversation closely and categorized them according to the following definitions. We define a *gap* (**GA**) as a silence at a speaker change (i.e. turn boundary, with speaker change from I-P or vice versa P-I) which is not perceived as unusually long. Following Sacks et al. (1978), a *lapse* (**LA**) is then distinguished from a gap by not only being longer by "rounds of possible self-selection" but also involving a discontinuity in the flow of conversation. More precisely, annotators were told to annotate a silence as a lapse for unusually long silences in communication between two individuals, at TRPs, and after which one participant (usually the interviewer in this dataset) initiates a new topic (topic shift). The final category, *attributable silence* (***AS***), occurs when the

[1] The python implementation used is at https://github.com/clp-research/deep_disfluency

TABLE 4 | The proposed interactional feature set.

Feature	Description
# LA	Total number of *LA* is sum of normalized no. of *LA* from *P–I* and *I-I*
Dur_LA	Sum of average *LA* duration from *P–I* and *I–I*
# GA	Total number of *GA* is the sum of normalized no. of *GA* from *P–I* and *I–P*
Dur_GA	Sum of average *GA* duration from *P–I* and *I–P*
# overlaps	No. of segments spoken simultaneously by both *P* and *I*. This feature indicates frequency of occurrence that may be attributed to speech initiation difficulties. (Young et al., 2016)
#Turn_switches per Minute	This is calculated by the number of turns per 60 s
Patient features	
# SP	Number of *SP* within *P* utterances normalized by the total # of words spoken by *P*
Dur_SP	Total duration of *SP* normalized by the total duration of speech by *P* without pauses
# LP	Number of *LP* within *P* utterances normalized by the total number of words spoken by *P*
Dur_LP	Total duration of *LP* normalized by the total duration of speech by *P* without pauses
# GA(P–I)	Number of *GA* at turn transition from *P–I* normalized by the total number of turns in the conversation
Dur_GA(P–I)	Average duration by considering the total duration of *GA (P–I)* divided by # *GA(P–I)*
# AS	Normalised number of attributable silence *AS* after posing the question from *I–P*
Dur_ AS	Average duration of *AS* from *I–P* with no response
Standardized pause rate (*SPR*)	*SPR* is obtained by the total number of words spoken by *P* divided by the sum of *SP* and *LP*.
Standardized phonation time (*SPT*)	*SPT* is the total number of words spoken by *P* to the total speech time of the patient excluding *SP* and *LP*.
Transformed phonation rate *TPR*	"The arcsine of the square root of the phonation rate (PR)" (Beltrami et al., 2018). *PR* is the speech time of *P* to the total speech time of *P* including *SP* and *LP*
Floor control ratio	This feature measures the relative amount of time (quantify dominance) the *P* spends speaking to the total speech time of the conversation (Aldeneh et al., 2019)
turn_length	This feature measures the number of words per turn spoken by *P*
speech_rate	Speech rate is the number of syllables per minute produced by *P*. It is calculated as the total numbers of syllables produced by *P* to the total speech time (in minutes)
Interviewers features	
# SP	Number of *SP* within *I* utterances normalized by the total # of words spoken by *I*
Dur_SP	Total duration of *SP* normalized by the total duration of speech by *I* without pauses
# LP	Number of *LP* within *I* utterances normalized by the # of words spoken by *I*
Dur_LP	Normalized duration of *LP*
# GA(I-P)	Number of *GA* at turn transition from *I–P* normalized by the total number of turns
Dur_GA(I–P)	Average duration of *GA (P–I)*
# LA(I–I)	Total # of *LA* is sum of all *LA (I–I)* normalized by # of turns
Dur_LA(I–I)	Average *LA* duration from *I–I* with the *topic shift*
# LA(P–I)	Normalized # of *LA* from *P–I* with a topic shift
Dur_LA(P–I)	Average *LA* duration from *P–I* with the *topic shift*
turn_length	This feature measures the # of words per turn spoken by *I*
speech_rate	This feature measures the number of syllable per minute during speech by *I*

current speaker selects another next speaker (by asking a question, by naming, or by looking at them), thereby putting the selected speaker under the obligation to speak next, but for one reason or another, that selected speaker does not respond; after the silence, the current speaker, therefore, continues the conversation (Elouakili, 2017). We define attributable silence as a longer silence after a question is asked from one party, no response from the other, and the first party then continues. Examples of these pause types with conversation samples are given in the Supplementary Materials. We also differentiated between speakers (patient *P* and interviewer *I*) by assigning speaker ID (*SP_ID*) to each labeled pause.

These annotations were performed using both transcripts and audio files using ELAN software (Sloetjes and Wittenburg, 2008).[2] To check the inter-rater agreement, two annotators annotated the silences of at least 0.5 s in one randomly selected AD patient dialogue; both had a good knowledge of linguistics and were familiar with the annotation rules. We use a multi-rater version of Cohen's κ (Cohen, 1960) as described by Siegel and Castellan (1988) to establish the agreement of annotators in terms of the overall agreement on all pause types, and also in terms of each pause type individually—see **Table 3**. We got an overall substantial agreement of $\kappa = 0.66$ for all categories of pauses. We got lower, though still moderately strong, κ values for *LP* and *SP* as these are pauses within the same speaker utterances and patients are older people with lower speech rates, making it more difficult to decide whether there is a relatively shorter or longer pause at certain lengths around the recommended boundary of 1.5 s.

Temporal Measures of Dialogue Interactions

Table 4 presents the extracted set of high-level interactional features to quantify the *P–I* interactions. There are 14 features for *P* and 12 features for *I* within the conversation and six features for overall conversation. This results in a set of 32 features representing the interaction within the natural dialogue conversations. We normalize the number of pauses within *P*

[2] https://archive.mpi.nl/tla/elan

TABLE 5 | Descriptive statistics (mean, SD) and statistical significance of the disfluency feature set. ** denotes highly significant at $p < 0.01$; * denotes significance at $p < 0.05$

Features	AD		Non-AD		Mann-Whitney *U* test	
	Mean	SD	Mean	SD	*P*	U
Patient features						
# *edit_terms*	0.029	0.009	0.017	0.006	**0.001****	183.5
# *Rpt*	0.027	0.015	0.011	0.13	**0.011***	172.0
# *Sub*	0.012	0.007	0.008	0.008	**0.045***	161.0
# *Del*	0.005	0.005	0.003	0.005	0.256	137.0
Interviewer features						
# *edit_terms*	0.009	0.011	0.004	0.004	**0.013***	170.5
# *Rpt*	0.010	0.008	0.007	0.006	**0.048***	157.0
# *Sub*	0.05	0.006	0.004	0.004	0.743	145.0
# *Del*	0.002	0.003	0.001	0.001	0.154	153.0

The boldfaced numbers indicate the best results.

or *I* by the number of words spoken by each respectively instead of normalizing by the number of utterances because it may be possible that when *P* speaks, they use a smaller number of words per utterance.

ANALYSIS AND EXPERIMENTS

Statistical Analysis

To investigate the importance of each feature, we calculated the mean and standard deviation (SD) for each group (AD and non-AD). We chose a nonparametric independent sample test (Mann-Whitney *U*) on disfluency and interactional features due to the small sample size. We applied a nonparametric test as a two-tailed test for unpaired samples and unequal variances. The value $p < 0.05$ was chosen for statistical significance. IBM SPSS version 26.0 was used for the statistical analysis.

Disfluency Features Analysis

Patient Features

Table 5 shows the results of our analysis indicating a significant difference between AD and non-AD patient groups in terms of the rate of patient edit terms, repeats, and substitution per word. The rate of edit terms is significantly higher ($p = 0.001$) for AD patients with a mean of 0.029 (SD = 0.009) compared to 0.017 (SD = 0.006) for non-AD patients. Furthermore, the rate of verbatim repeat disfluencies is significant ($p = 0.011$) with a higher mean value for AD patients than non-AD patients (0.027 vs. 0.011). The findings also indicate a significant correlation between conditions and substitution disfluencies ($p = 0.045$), again with higher rates for AD patients vs. non-AD patients (0.012 vs. 0.008). Disfluencies are known to be symptomatic of communication difficulties. People who suffer from AD typically experience communication problems through weak conversation flow; it is reasonable that this will be observable through increased disfluencies in dialogue. The rate of delete disfluencies is, however, not found to be significantly different between AD and non-AD patients, possibly due to lack of data as they are very rare.

Interviewer Features

As with patient features, we found that there is a significantly greater rate of edit terms in conversations with AD patients ($p = 0.013$) with a mean value of 0.009 (SD = 0.011) compared to 0.004 (SD = 0.004) for those with non-AD patients. The rate of repeat disfluencies ($p = 0.048$) is also significantly greater with a mean value of 0.010 (SD = 0.008) in interviewer speech with AD patients and a mean value of 0.007 (SD = 0.006) in interviewer speech with non-AD individuals. The rate of delete and substitution disfluencies are not found to be significantly different in interviewer speech with AD and non-AD patients. The fact that there are more disfluencies in the interviewer's speech suggests that trouble with communication is shared between both participants, in line with the Conversation Analytic emphasis on collaborative achievement.

Interactional Features Analysis

Table 6 presents the mean, SD, the *p*-values, and test statistic U (for Mann-Whitney *U* test) for each of the interactional features reported in **Table 4**. Significant differences between the AD and non-AD groups are marked in bold. Overall, the total number of *GA* and the total number of *LA* are found to be significantly higher in the AD group. There were fewer turn switches in AD dialogues with a mean of 2.544 compared to non-AD dialogues with a higher mean of 3.510. **Figure 1** shows the distributions of three significant features with **Figure 1A–C** and **Figure 1D** representing the distribution of a nonsignificant feature, i.e. average duration of *LA (P–I)* between AD and non-AD groups. There is a great number of *AS* shown in **Figure 1A** with longer silences in the AD group than the non-AD group. The Y-axis shows the normalized duration while the X-axis shows the frequency of duration of the *AS* in each group.

Patient Features

Our analysis found that the patient's long pauses, duration of long pause, number of gaps from *P–I*, and duration of *AS* exhibit significant differences between AD and non-AD patient groups. Standardized phonation time of patients is significantly lower for AD patients, with a mean of 2.113 and variability of 0.531 for AD patients, and a mean of 2.839 for non-AD patients. Mean turn length is significantly higher at 22.52 s for non-AD patients compared to 12.142 for AD patients. These results suggest AD patients produce a greater number of pauses with a longer duration (>1.5 s), with slower speech rates than non-AD patients. These longer pauses within the patients' utterances signal the difficulty in lexical search and semantic processing problems of finding key components related to events, places, etc. Additionally, the results suggest that AD patients exhibit higher variability in the time they either respond to questions by clinicians (resulting in high values for the number of gaps from *I–P* with larger delays) or they preferred attributable silences (mean duration of 2.468 for AD patients as compared to 0.414 for non-AD patients) instead of response. Notably, the floor control ratio is higher for non-AD patients, suggesting that AD patients hold the floor for less time compared to non-AD patients. The number of short pauses and duration of short

TABLE 6 | Descriptive statistics (mean, SD) and statistical significance for our interactional feature set. We report *p* values obtained from Mann-Whitney U tests against a null hypothesis with no differences in distributions of these interactions on AD. ** denotes highly significant at *p* < 0.01; * denotes significance at; - shows a trend toward significance at *p* < 0.1.

Features	**AD**		**Non-AD**		**Mann-Whitney *U* test**	
	Mean	**SD**	**Mean**	**SD**	***p***	**U**
#LA	0.051	0.053	0.011	0.020	**0.013***	171.5
Dur_LA	3.195	2.592	1.041	1.927	**0.026***	166.0
# GA	0.228	0.121	0.104	0.071	**0.010***	174.0
Dur_GA	1.400	0.464	1.100	0.245	0.067-	156.0
# overlaps	0.073	0.029	0.109	0.082	0.595	99.0
#Turn_switches per Minute	2.544	0.835	3.510	1.447	**0.026***	59.5
Patient features						
# SP	0.034	0.013	0.032	0.018	0.455	130.5
Dur_SP	0.064	0.022	0.082	0.06	0.254	85.0
# LP	0.022	0.016	0.012	0.017	**0.013***	171.5
Dur_LP	0.106	0.078	0.054	0.065	**0.016***	169.5
# GA(P–I)	0.103	0.067	0.052	0.054	**0.015***	170.5
Dur_GA(P–I)	1.515	0.820	1.000	0.368	0.098-	152.5
# AS	0.010	0.013	0.002	0.002	0.067-	157.0
Dur_ AS	2.468	3.243	0.414	0.724	**0.037***	163.0
(SPR)	22.158	12.54	36.40	28.19	0.137	76.0
(SPT)	2.113	0.531	2.839	0.060	**0.002****	41.0
TPR	1.041	0.115	1.114	0.157	0.081-	70.0
Floor control ratio	0.596	0.172	0.712	0.183	0.098-	72.5
turn_length	12.142	6.59	22.52	20.34	**0.007****	168.5
speech_rate	164.91	35.74	180.1	37.82	0.345	89.0
Interviewer features						
# SP	0.013	0.009	0.017	0.02	0.935	110.0
Dur_SP	0.029	0.020	0.034	0.036	0.902	109.0
# LP	0.006	0.006	0.005	0.007	0.126	149.5
Dur_LP	0.033	0.023	0.021	0.037	0.061-	157.5
# GA(I–P)	0.125	0.068	0.052	0.033	**0.002****	184.5
Dur_GA(I–P)	1.363	0.365	1.011	0.301	**0.041***	161.5
# LA(I–I)	0.020	0.023	0.027	0.068	0.305	137.5
Dur_LA(I–I)	3.291	3.696	1.316	1.951	0.106	151.5
# LA(P–I)	0.031	0.037	0.002	0.003	**0.009****	175.0
Dur_LA(P–I)	2.552	2.161	1.163	2.317	0.081-	155.0
turn_length	9.155	4.320	23.31	22.31	**0.001***	34.0
speech_rate	195.49	32.89	183.05	43.09	0.325	137.0

The boldfaced numbers indicate the best results.

pauses were not found to be significant between AD and non-AD patients, suggesting that short pauses are present naturally for breathing and for planning at the word or phrase level.

Interviewer Features

We found that the duration of *LP* is approaching significance with the mean 0.033 (SD = 0.023) for interviewers with an AD patient being higher than 0.021 (SD = 0.037) for those with non-AD patients. While only a tendency, we can tentatively conclude interviewers tend to insert longer silences while interacting with AD patients. The number of *GA* at *I–P* turn changes is significantly greater at turn exchanges with AD patients, with an average of 0.103 with a longer duration of 1.515 compared to the mean of 0.052 with a relatively shorter duration on average of 1.011 at turn exchanges with non-AD patients. The number of *LA* is also highly predictive among the two groups in the *P–I* turn changes. This means that the frequency of initiating a new topic by the interviewer after a considerable amount of silence after the patient has stopped speaking is higher in the AD group with a mean of 0.031, compared to 0.002 for non-AD patients. Finally, we found that the average turn length of interviewers interacting with AD patients is 9.155 s (SD = 4.320) compared to 23.31 s (SD = 22.31) with non-AD interactions, the mirror image of the case with patient turn length, where AD patients have far longer turns. This reveals that although the interviewers paused for longer periods within their turns while interacting with AD patients they also tend to speak for a shorter period of time.

Our study provides strong evidence that these interactional features including pause duration, gaps, lapse duration, presence of attributable silences, phonation time, and turn length seem to be sensitive markers of cognitive decline and also distinguish the AD group from the non-AD group.

Classification Experiments

Our final goal is to perform a classification task to assess whether AD prediction can be improved by integrating these inter-speaker interactional features with the intra-speaker disfluency features. We study the influence of these features using three machine

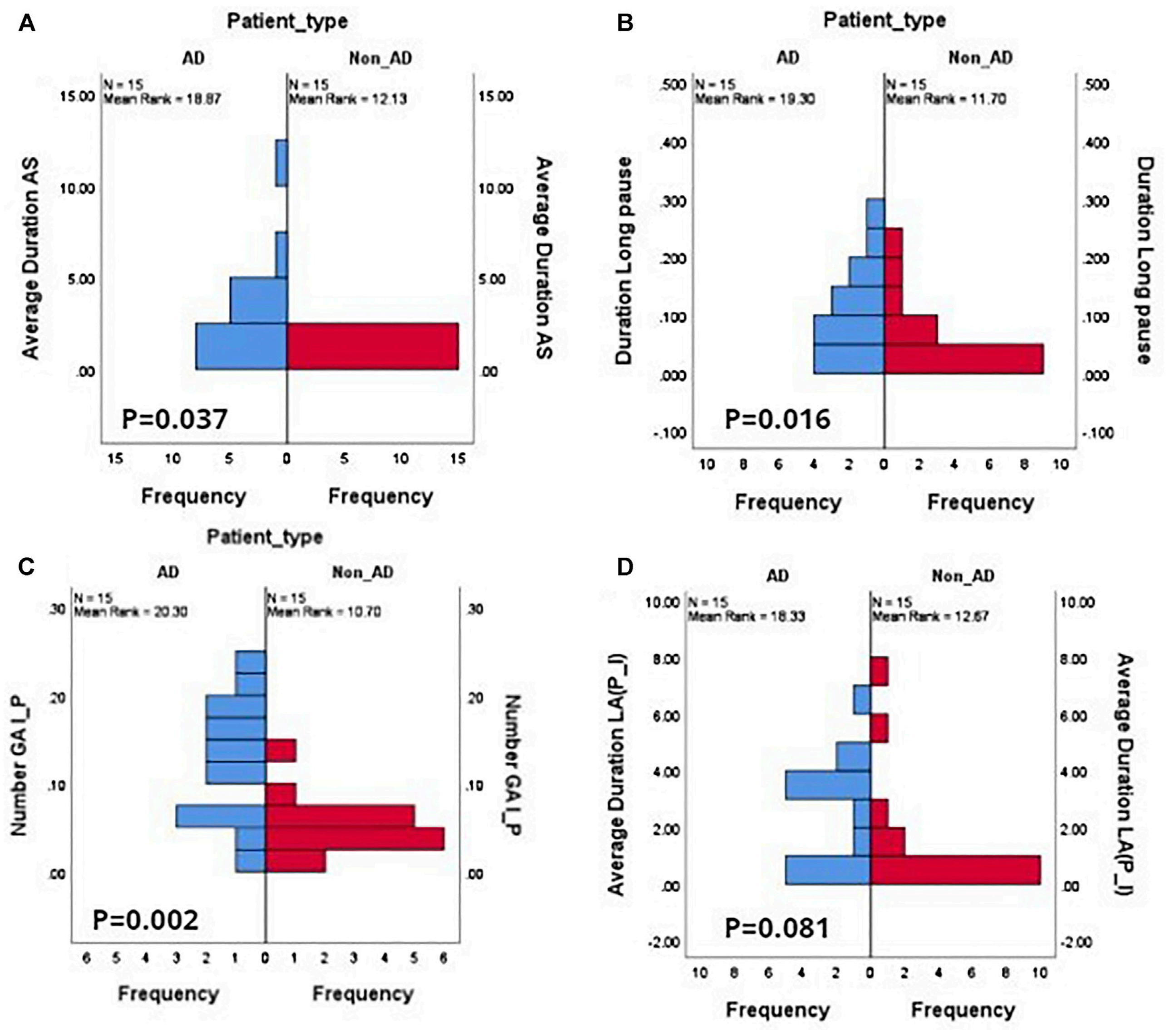


FIGURE 1 | Feature value histograms for a selection of different pause types, showing differences in distributions between AD and non-AD dialogues. **(A)** Average duration of patient attributable silences *AS*; **(B)** duration of patient long pauses *LP*; **(C)** frequency distribution of interviewer-to-patient gaps *GA(I–P)*; **(D)** duration of patient-to-interviewer lapses *LA(P–I)*. **(A)**, **(B)**, **(C)** show distributions that are significant at $p < 0.05$, while for **(D)** $0.05 < p < 0.1$.

learning classifiers: logistic regression (LR), support vector machines (SVM), and multilayer perceptron (MLP). We train each classifier using disfluency features, interactional features, and then by combining both. As the dataset is fairly small, we did not use separate splits of data for train and test, but rather follow a leave-one-out cross validation (LOOCV) scheme to get a better estimation of generalization accuracy. This process involves selecting one participant as a test and training the classifier on the remaining instances. This process is repeated until all instances have been selected for testing. The resulting accuracies on all folds are then aggregated into a final score. We build our models using the Scikit-Learn library (Pedregosa et al., 2011). We optimize our models with the following hyper-parameters: logistic regression with $C \in \{0.001,0.01,0.1,1,10,100,1000\}$ using the "liblinear" solver; SVM with $C \in \{$ 0.1, 1, 10, 100, 1000$\}$, $\gamma \in$ $\{$1, 0.1, 0.01, 0.001, 0.0001$\}$, using the kernels "rbf" and "poly"; and MLP with the "relu" activation function, hidden layer sizes of (2,3), and (3,4) and an initial learning rate of 0.01. We also performed a recursive feature elimination (RFE) method on both interactional and disfluency feature set to eliminate the weakest features with the purpose of removing any dependencies and colinearity. RFE is a feature selection method that removes a certain number of weak features per iteration and fits the model with the remaining features. We then train each classifier with the top 15 ranked features based on RFE.

TABLE 7 | Comparison of results for the AD classification with three classifiers with LOOCV.

Model	Feature set	Accuracy	Precision	Recall	F1 score	AUC
LR	Language	0.77	0.75	0.80	0.77	0.74
	Dialogue	0.80	0.81	0.80	0.80	0.80
	Both	0.87	0.87	0.87	0.87	0.84
	RFE (15)	0.83	0.86	0.80	0.83	0.81
SVM	Language	0.83	0.83	0.83	0.83	0.85
	Dialogue	0.83	0.83	0.83	0.83	0.87
	Both	**0.90**	**0.90**	**0.90**	**0.90**	**0.89**
	RFE (15)	0.87	0.87	0.87	0.87	0.85
MLP	Language	0.77	0.75	0.80	0.77	0.75
	Dialogue	0.80	0.77	0.76	0.76	0.79
	Both	0.80	0.80	0.80	0.80	0.81
	RFE (15)	0.80	0.80	0.80	0.80	0.80

The boldfaced numbers indicate the best results.

Because our dataset is balanced, we reported our results in terms of accuracy, precision, recall, F1 score, and area under the ROC curve (AUC) as evaluation metrics. Precision measures what percentage of AD predictions correspond to real cases of AD (i.e. true positive divided by true positive and false positive). Recall measures the percentage of the actual AD occurrences that were detected (i.e. true positives divided by false negative plus true positive). F1 is the harmonic mean of precision and recall. AUC is commonly used for evaluating the performance of clinical diagnostic and predictive models (Zou et al., 2007). The ROC curve is used to show the trade-off between true positive rate (TPR, recall of the AD class) and false positive rate (FPR, one-recall of the non-AD class). Different clinical diagnostic scenarios may call for different TPR/FPR trade-offs, so the area under the curve (AUC) is used to express the overall level of diagnostic power; AUC greater than 0.75 is usually recommended for clinical purposes (Orimaye et al., 2017).

Classification Results and Discussion

Table 7 provides the classification accuracy measures obtained using an individual group of features for combining both sets of features and when applying RFE top 15 selected features against all three classifier algorithms—LR, SVM, and MLP. It can be seen that the SVM outperformed both LR and MLP using disfluency features, interactional features, the combination of both, and with RFE-based top 15 features. Comparing the two feature sets, the best scores attained (with the SVM) are in fact identical with accuracies of 83%. However, by combining the two feature sets we achieved the highest accuracy of 90% with an F1 score of 0.90 with the SVM classifier. With LR, we achieved an accuracy of 77% with disfluency features, 80% with interactional features, and an increase in accuracy of roughly 7% when combining both feature sets with 87%.

MLP performed similarly to LR for disfluency features, with the same accuracy and F1 score; however, it performs slightly worse with the interactional features with an F1 score of 0.76 compared to LR and SVM. The combination of both feature sets showed an increase in the F1 score to 0.80. From the overall accuracy results with MLP, we can draw the conclusion that as MLP is a feed-forward neural network with more parameters and is a more data-hungry algorithm, the small number of samples and small feature space available for training is suboptimal.

Luz et al. (2018) used a probabilistic graphical model to classify AD patients in the CCC, using a slightly bigger dataset but with shorter dialogue conversations. They used only interactional features, and achieved comparable accuracies of 0.757 with LR and 0.837 with SVM classifiers; but did not investigate the role of different pause types, or the combination with fluency. Interestingly, they found that AD patients produce longer turns with more words and a higher speech rate; this contrasts with our results, in which AD patients produce fewer words than non-AD patients, with lower speech rates. We note that our findings align better with other research (Martínez-Sánchez et al., 2013; Kavé and Dassa, 2018; Pistono et al., 2019a; Themistocleous et al., 2020). Mirheidari et al. (2019) went a step further, combining CA-inspired interaction features including turn-taking behavior with some acoustic and language features, to achieve a classification accuracy of 90% similar to this study. However their approach is based on structured interviews with chosen topics and question types, in more clinical settings, and the use of features that directly target particular aspects of this structure (e.g. responses to particular setting-specific questions).

Effect of Disfluency Features

We found that disfluency tags help as features in AD detection. With these disfluency features, we got the highest accuracy of 83% with the SVM classifier, an identical accuracy to using interactional features. It is also worth examining the ROC AUC as it evaluates the different classifiers at different true positive rates and false positive rates. **Figure 2A** shows the ROC curve for the disfluency features with the SVM, with AUC 0.85, and with TPR 0.87 and FPR 0.20 at the chosen trade-off point. We have chosen this trade-off point as it gives maximum accuracy.

Effect of Interactional Features

Our interactional features produced promising results in distinguishing AD from non-AD with overall accuracy reaching 83% with the SVM classifier, showing that interactional patterns can provide salient cues to the detection of AD in dialogues. The results are further enhanced when adding with disfluency language feature reaching an accuracy of 90% and F1 score of 0.90. These results suggest that different pauses behavior not only indicate word-finding difficulties as AD progresses but also mark disfluency—in certain situations showing these were used to sustain social interaction as part of compensatory language (e.g. in the case of attributable silences). The corresponding ROC curve is shown in **Figure 2B** with AUC 0.87, and the chosen trade-off between TPR and FPR (0.80 vs 0.13). It can also be seen in **Figure 2C** that combining these interactional features with language features over dialogues had the effect of improving classification performance overall to AUC = 0.89, and improving trade-offs between true positive (0.93) and false positive rates (0.13), reducing the false positives while increasing the true positives.

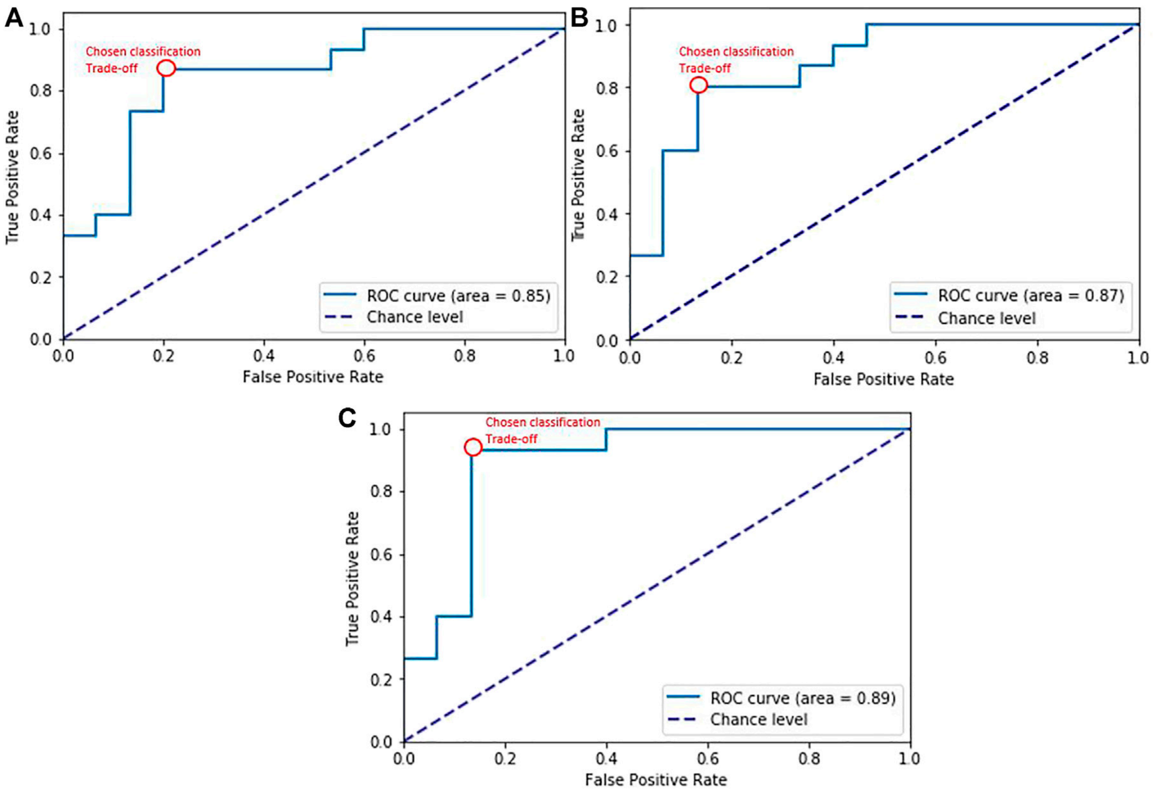


FIGURE 2 | ROC curves for SVM classification experiments with **(A)** disfluency features, **(B)** interactional features, **(C)** the combined feature set. The red bubble shows the chosen trade-off point for the classification experiment results in **Table 7**.

TABLE 8 | Top 15 ranked features including disfluency and interactional features by RFE.

Features	Type	Ranking
Dur_AS	Interactional	1
turn_switches_per_minute	Interactional	2
Dur_LA	Interactional	3
Dur_LA (P-I)	Interactional	4
#GA	Interactional	5
TPR	Interactional	6
P_RPT	Language	7
I_turn_length	Interactional	8
Dur_LA (I-I)	Interactional	9
# LA	Interactional	10
I_edit_terms	Language	11
P_edit_terms	Language	12
SPT	Interactional	13
P_Turn_Length	Interactional	14
I_Speech_rate	Interactional	15

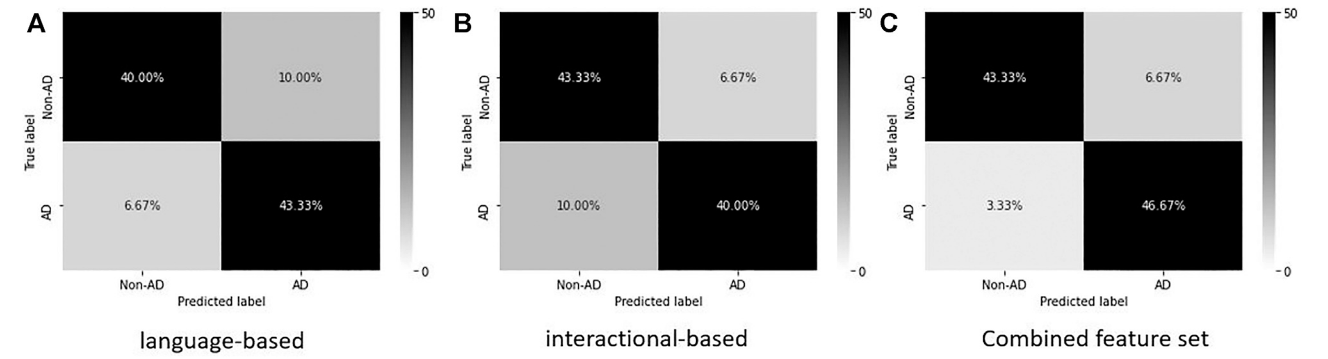
We also reported the top 15 ranked features based on RFE as shown in **Table 8**. These features were also found to be significant in our statistical analysis (see **Table 6**). As with the statistical test-based features, *Dur_AS* has been picked and is ranked first as the most significant. This confirms the findings of Levinson (1983) concerning attributable silences and aligns with conversation analysis studies showing that individuals with cognitive decline resort to silence rather than other means of communication to avoid giving a dispreferred response. Among the other useful features, not only the number of gaps and lapses are found to be important but also the duration of gaps and lapses are observed differently in both groups. Turn switches per minute, patient turn lengths, and standardized phonation time are negatively correlated with AD patients with higher mean values for non-AD. That means turn switches happen more frequently, with longer turn lengths, in conversations with non-AD patients compared to AD individuals.

Error Analysis

The results in **Table 9** show that the SVM model with disfluency and interactional features attained the highest F1 score, precision, and recall for both AD and non-AD classes; we show both classes to provide a measure of both sensitivity (recall of the positive AD class) and specificity (recall of the non-AD class), standard measures for diagnostic tests. Note that due to the small dataset, differences between modes are indicative rather than statistically significant—see the confidence intervals in **Table 9**. The model achieves F1 scores of 0.90 for both the AD and the non-AD classes. Combining the disfluency features with interactional features particularly improves the recall of the AD class (i.e. improves the sensitivity of the classifier): the SVM model with both feature sets has a recall of 0.93, improving overused disfluency features alone at 0.87 and over

TABLE 9 | Results of AD classification task with SVM classifiers with different feature sets, using LOOCV, with 95% confidence intervals (CI).

Model	Class	Precision	Recall	F1 score	Accuracy	95% CI
SVM	AD	0.81	0.87	0.84	0.83	0.70–0.96
(Language)	Non-AD	0.86	0.80	0.83	–	–
SVM	AD	0.86	0.80	0.83	0.83	0.70–0.96
(Dialogue)	Non-AD	0.81	0.87	0.84	–	–
SVM	AD	0.87	0.93	0.90	0.90	0.79–0.99
(Both)	Non-AD	0.93	0.87	0.90	–	–



language-based

interactional-based

Combined feature set

FIGURE 3 | Confusion matrices for AD classification task with different feature sets.

the 0.80 achieved with interactional features. The specificity (recall for the non-AD class) was lowest when using language features only at 0.80, significantly lower than the 0.87 achieved by both using dialogue features alone and combining both feature sets. A balanced F1 score for both the AD and non-AD classes with all three combinations was achieved overall with our chosen threshold (0.84 vs 0.83 for disfluency features, 0.83 vs 0.84 with interactional features, and 0.90 for the combined feature sets). Depending on the application the model is used for, higher sensitivity or higher specificity for AD detection will be more or less desirable and this can be achieved in line with the AUC results shown in **Figure 2**, but as it stands using the combined feature set considerably increases the sensitivity of AD diagnosis over the most sensitive single feature set classifier (language features) while maintaining a high specificity on par with that achieved using dialogue features. We can observe the confusion matrices of predictions of the SVM Model with language, interactional, and combining both in **Figure 3** which show the influence of (A) and (B) on (C).

CONCLUSION

This study investigated techniques for the diagnosis of dementia using features of disfluency and interaction in natural dialogue conversation, rather than relying on linguistic features alone, or either structured interviews or picture description tasks. We first performed a statistical analysis on the disfluency and interactional features. This analysis indicates that the relative frequency of edit terms, verbatim repeats, and substitution disfluencies are derived measures of disfluency in natural conversations that have different distributions in interviews with AD patients and those with non-AD patients. We also found that most of the interactional features, including attributable silences, gaps, lapses, turn lengths, and turn switches per minute, are sensitive cues in discriminating AD patients from non-AD patients. We also observed that in natural conversation not only are patients' conversation characteristics affected but also distinctive patterns can be observed in interviewers' or carers' conversational behavior when talking to AD patients.

Our results showed the efficacy of detecting AD from dialogue using machine learning classifiers with different feature sets, which involved using them separately and then combining them. We obtained identical overall accuracy scores when both using disfluency features and interactional features separately at 83%. Disfluency features hold predictive power for the identification of AD, giving rise to a classifier with higher sensitivity (recall on AD = 0.87 vs 0.80), while the interactional dialogue features allow a higher specificity of AD detection (recall of non-AD = 0.87 vs 0.80). However combining the linguistic and interactional features obtained the most sensitive and specific automatic diagnostic classifier (recall on AD = 0.93, recall on non-AD = 0.87) with an overall accuracy of 90% on a balanced dataset, suggesting the potential benefits of integrating these features into clinical assessments via natural conversation as diagnostics.

We further plan to extend this study by introducing language markers associated with AD severity beyond disfluencies, as well as interactions between them. In particular, we want to use a more principled approach to lexical markers and measures of grammatical fluency. We also plan to use acoustic features, including prosodic, voice quality, and spectral features, which contribute to AD recognition and have higher correlations and interact with linguistic information. At the interactional feature level, we plan to include dialogue act (DA) tags that provide more of the speaker's illocutionary content at the utterance level, including different tags for questions, answers types, clarification requests, signals of misunderstanding, and then use sequences of these DA tags to predict the disrupted communication patterns in natural conversations with AD patients.

While the results are promising, there are limitations to the data used in this study. The CCC only contains older patients with diagnosed dementia at moderate stages, so it can only allow us to observe the patterns associated with AD at a relatively advanced stage, and not whether these extend to early-stage diagnosis. To overcome this, we need to collect new datasets that contain spontaneous speech conversations with patients at different stages of dementia to analyze disfluencies and interactional features shown in early cognitive decline.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Due to privacy concerns of patient's data, data is not publically available and was accessed after Ethical research approval (QMERC2019/04) in the present study. Requests to access these datasets should be directed to https://carolinaconversations.musc.edu/help/access.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Queen Mary Ethics of Research Committee, Queen Mary University of London, and the Medical University of South Carolina (MUSC). All subjects provided written informed consent in the original study by the MUSC. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SN contributed to the design of the study, interpreted the data, took part in annotation protocol design, and performed statistical analysis on interactional features. MR performed disfluency feature analysis and performed classification experiments by combining disfluency and interactional features and helped in drafting the manuscript. MP contributed to the interpretation of data, helped in preparing the annotation protocol, supervised the whole process from annotation to statistical analysis to experimentation, and revised the manuscript critically. JH contributed to the interpretation of data, calculated the kappa agreement for pauses types, and revised the manuscript critically. All authors have contributed to this study and gave final approval for this manuscript and agree to be accountable for the content of the work.

FUNDING

MP was partially supported by the EPSRC under grant EP/S033564/1 and by the European Union's Horizon 2020 programme under grant agreements 769661 (SAAM, Supporting Active Ageing through Multimodal coaching) and 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.

ACKNOWLEDGMENTS

The contribution of Jorge Del-Bosque-Trevino to the annotation protocol design process and the annotations themselves is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2021.640669/full#supplementary-material

REFERENCES

Abel, S., Huber, W., and Dell, G. S. (2009). Connectionist Diagnosis of Lexical Disorders in Aphasia. *Aphasiology* 23, 1353–1378. doi:10.1080/02687030903022203

Addlesee, A., Eshghi, A., and Konstas, I. (2019). Current Challenges in Spoken Dialogue Systems and Why They Are Critical for Those Living with Dementia. arXiv preprint arXiv:1909.06644

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected Speech as a Marker of Disease Progression in Autopsy-Proven Alzheimer's Disease. *Brain a J. Neurol.* 136, 3727–3737. doi:10.1093/brain/awt269

Aldeneh, Z., Jaiswal, M., Picheny, M., McInnis, M., and Provost, E. M. (2019). Identifying Mood Episodes Using Dialogue Features from Clinical Interviews. arXiv preprint arXiv:1910.05115doi:10.21437/interspeech.2019-1878

Bayles, K. A., and Boone, D. R. (1982). The Potential of Language Tasks for Identifying Senile Dementia. *J. Speech Hear. Disord.* 47, 210–217. doi:10.1044/jshd.4702.210

Becker, J. T., Boller, F., Lopez, O., Saxton, J., and McGonigle, K. (1994). The Natural History of Alzheimer's Disease. *Arch. Neurol.* 51, 585–594. doi:10.1001/archneur.1994.00540180063015

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., and Calzà, L. (2018). Speech Analysis by Natural Language Processing Techniques: a Possible Tool for Very Early Detection of Cognitive Decline? *Front. Aging Neurosci.* 10, 369. doi:10.3389/fnagi.2018.00369

Brandt, J. (1991). The Hopkins Verbal Learning Test: Development of a New Memory Test with Six Equivalent Forms. *Clin. Neuropsychologist* 5, 125–142. doi:10.1080/13854049108403297

Broderick, B. M., Tou, S. L., and Provost, E. M. (2018). TD-P-014: Cogid: A Speech Recognition Tool for Early Detection of Alzheimer's Disease. *Alzheimer's Demen.* 14, P191–P192. doi:10.1016/j.jalz.2018.06.2030

Chapman, S. B., Zientz, J., Weiner, M., Rosenberg, R., Frawley, W., and Burns, M. H. (2002). Discourse Changes in Early Alzheimer Disease, Mild Cognitive Impairment, and normal Aging. *Alzheimer Dis. Associated Disord.* 16, 177–186. doi:10.1097/00002093-200207000-00008

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* 20, 37–46. doi:10.1177/001316446002000104

Crystal, D., and Davy, D. (2016). *Investigating English Style*. (London: Routledge). doi:10.4324/9781315538419

Davis, B. H., and Maclagan, M. (2009). Examining Pauses in Alzheimer's Discourse. *Am. J. Alzheimers Dis. Other Demen.* 24, 141–154. doi:10.1177/1533317508328138

Davis, B. H., and Maclagan, M. (2010). Pauses, Fillers, Placeholders and Formulaicity in Alzheimer's Discourse. *Fillers, Pauses and placeholders* 93, 189–216. doi:10.1075/tsl.93.09dav

Davis, B., Maclagan, M., and D, S. (2014). "Exploring Interactions between Visitors and Residents with Dementia, with a Focus on Questions and the Responses They Evoke," in *The Routledge Handbook of Language and Health Communication* (The city: Routledge), 344–360.

de la Fuente Garcia, S., Ritchie, C. W., and Luz, S. (2019). Protocol for a Conversation-Based Analysis Study: Prevent-Ed Investigates Dialogue Features that May Help Predict Dementia Onset in Later Life. *BMJ open* 9, e026254. doi:10.1136/bmjopen-2018-026254

de Lira, J. O., Ortiz, K. Z., Campanha, A. C., Bertolucci, P. H. F., and Minett, T. S. C. (2011). Microlinguistic Aspects of the Oral Narrative in Patients with Alzheimer's Disease. *Int. Psychogeriatr.* 23, 404–412. doi:10.1017/s1041610210001092

Elouakili, S. (2017). A Conversation Analysis Approach to Attributable Silence in Moroccan Conversation. *Ire* 5, 1–21. doi:10.5296/ire.v5i2.11369

Elsey, C., Drew, P., Jones, D., Blackburn, D., Wakefield, S., Harkness, K., et al. (2015). Towards Diagnostic Conversational Profiles of Patients Presenting with Dementia or Functional Memory Disorders to Memory Clinics. *Patient Edu. Couns.* 98, 1071–1077. doi:10.1016/j.pec.2015.05.021

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental State". *J. Psychiatr. Res.* 12, 189–198. doi:10.1016/0022-3956(75)90026-6

Forbes-McKay, K. E., and Venneri, A. (2005). Detecting Subtle Spontaneous Language Decline in Early Alzheimer's Disease with a Picture Description Task. *Neurol. Sci.* 26, 243–254. doi:10.1007/s10072-005-0467-9

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016a). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *J. Alzheimers Dis.* 49, 407–422. doi:10.3233/JAD-150520

Fraser, K. C., Rudzicz, F., and Hirst, G. (2016b). "Detecting Late-Life Depression in Alzheimer's Disease through Analysis of Speech and Language," in *Proc. CLPsych* (San Diego, CA, USA: Association for Computational Linguistics)), 1–11.

Gayraud, F., Lee, H.-R., and Barkat-Defradas, M. (2011). Syntactic and Lexical Context of Pauses and Hesitations in the Discourse of Alzheimer Patients and Healthy Elderly Subjects. *Clin. Linguistics Phonetics* 25, 198–209. doi:10.3109/02699206.2010.521612

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). "Switchboard: Telephone Speech Corpus for Research and Development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on* (IEEE Computer Society)), 1, 517–520.

Goodglass, H., Kaplan, E., Weintraub, S., and Barresi, B. (2001). *The boston Diagnostic Aphasia Examination*. (Philadelphia, PA: Lippincott, Williams & Wilkins)

Hamilton, H. E. (2005). *Conversations with an Alzheimer's Patient: An Interactional Sociolinguistic Study*. New York: Cambridge University Press.

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., and Kálmán, J. (2010). Temporal Parameters of Spontaneous Speech in Alzheimer's Disease. *Int. J. speech-language Pathol.* 12, 29–34. doi:10.3109/17549500903137256

Hough, J., and Schlangen, D. (2017). "Joint, Incremental Disfluency Detection and Utterance Segmentation from Speech," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 1, 326–336.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., et al. (2014). "Aided Diagnosis of Dementia Type through Computer-Based Analysis of Spontaneous Speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 27–37.

Jones, D., Drew, P., Elsey, C., Blackburn, D., Wakefield, S., Harkness, K., et al. (2016). Conversational Assessment in Memory Clinic Encounters: Interactional Profiling for Differentiating Dementia from Functional Memory Disorders. *Aging Ment. Health* 20, 500–509. doi:10.1080/13607863.2015.1021753

Kalbe, E., Kessler, J., Calabrese, P., Smith, R., Passmore, A. P., Brand, M., et al. (2004). Demtect: a New, Sensitive Cognitive Screening Test to Support the Diagnosis of Mild Cognitive Impairment and Early Dementia. *Int. J. Geriat. Psychiatry* 19, 136–143. doi:10.1002/gps.1042

Kavé, G., and Dassa, A. (2018). Severity of Alzheimer's Disease and Language Features in Picture Descriptions. *Aphasiology* 32, 27–40. doi:10.1080/02687038.2017.1303441

Kavé, G., and Levy, Y. (2003). Morphology in Picture Descriptions provided by Persons with Alzheimer's Disease. *J. Speech, Lang. Hearing Res.* 46 (2), 52–341.

Kopelman, M., Wilson, B., and Baddeley, A. (1990). *The Autobiographical Memory Interview (Manual)*. Bury St. Edmunds, England: Thames Valley Test Company.

Le Boeuf, C. (1976). *Raconte. . .: 55 historiettes en images*. L'école.

Levelt, W. (1983). Monitoring and Self-Repair in Speech. *Cognition* 14, 41–104. doi:10.1016/0010-0277(83)90026-4

Levinson, S. C. (1983). *Pragmatics*. Cambridge UK: Cambridge University Press. doi:10.1017/cbo9780511813313

López-de-Ipiña, K., Alonso, J.-B., Travieso, C., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., et al. (2013). On the Selection of Non-invasive Methods Based on Speech Analysis Oriented to Automatic Alzheimer Disease Diagnosis. *Sensors* 13, 6730–6745. doi:10.3390/s130506730

Luz, S., de la Fuente, S., and Albert, P. (2018). "A Method for Analysis of Patient Speech in Dialogue for Dementia Detection," in *Proceedings of the LREC 2018 Workshop Resources and Processing of Linguistic, Para-Linguistic and Extra-linguistic Data from People with Various Forms of Cognitive/psychiatric Impairments (RaPID-2)*. Editor D. Kokkinakis, 25–42.

Martínez-Sánchez, F., Meilán, J. J. G., García-Sevilla, J., Carro, J., and Arana, J. M. (2013). Oral reading Fluency Analysis in Patients with Alzheimer Disease and Asymptomatic Control Subjects. *Neurología (English Edition)* 28, 325–331. doi:10.1016/j.nrleng.2012.07.017

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical Diagnosis of Alzheimer's Disease: Report of the NINCDS-ADRDA Work Group* under the Auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 34, 939. doi:10.1212/wnl.34.7.939

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Jr, Kawas, C. H., et al. (2011). The Diagnosis of Dementia Due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease. *Alzheimer's Demen.* 7, 263–269. doi:10.1016/j.jalz.2011.03.005

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019). Dementia Detection Using Automatic Analysis of Conversations. *Comp. Speech Lang.* 53, 65–79. doi:10.1016/j.csl.2018.07.006

Nasreen, S., Purver, M., and Hough, J. (2019). "A Corpus Study on Questions, Responses and Misunderstanding Signals in Conversations with Alzheimer's Patients," in *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue-Full Papers* (London, United Kingdom: SEMDIAL), 13, 89–98. *http://semdial.org/anthology/Z19-Nasreen semdial*.

Noone, P. (2015). Addenbrooke's Cognitive Examination-III. *Occmed* 65, 418–420. doi:10.1093/occmed/kqv041

Orimaye, S. O., Wong, J. S., Golden, K. J., Wong, C. P., and Soyiri, I. N. (2017). Predicting Probable Alzheimer's Disease Using Linguistic Deficits and Biomarkers. *BMC Bioinformatics* 18, 34. doi:10.1186/s12859-016-1456-0

Pasquier, F., Lebert, F., Grymonprez, L., and Petit, H. (1995). Verbal Fluency in Dementia of Frontal Lobe Type and Dementia of Alzheimer Type. *J. Neurol. Neurosurg. Psychiatry* 58, 81–84. doi:10.1136/jnnp.58.1.81

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in python. *J. Machine Learn. Res.* 12, 2825–2830.

Perkins, L., Whitworth, A., and Lesser, R. (1998). Conversing in Dementia: A Conversation Analytic Approach. *J. Neurolinguist.* 11, 33–53. doi:10.1016/s0911-6044(98)00004-9

Pistono, A., Jucla, M., Bézy, C., Lemesle, B., Men, J., and Pariente, J. (2019a). Discourse Macrolinguistic Impairment as a Marker of Linguistic and Extralinguistic Functions Decline in Early Alzheimer's Disease. *Int. J. Lang. Commun. Disord.* 54, 390–400. doi:10.1111/1460-6984.12444

Pistono, A., Pariente, J., Bézy, C., Lemesle, B., Le Men, J., and Jucla, M. (2019b). What Happens when Nothing Happens? an Investigation of Pauses as a Compensatory Mechanism in Early Alzheimer's Disease. *Neuropsychologia* 124, 133–143. doi:10.1016/j.neuropsychologia.2018.12.018

Pope, C., and Davis, B. H. (2011). Finding a Balance: The Carolinas Conversation Collection. *Corpus Linguistics Linguistic Theor.* 7, 143–161. doi:10.1515/cllt.2011.007

Rohanian, M., Hough, J., and Purver, M. (2020). Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech. *Proc. Interspeech* 2020, 2187–2191. doi:10.21437/Interspeech.2020-2721

Rohanian, M., and Hough, J. (2020). "Re-framing Incremental Deep Language Models for Dialogue Processing with Multi-Task Learning," in *Proceedings of the 28th International Conference on Computational Linguistics*, 497–507.

Rosen, W. G., Mohs, R. C., and Davis, K. L. (1984). A New Rating Scale for Alzheimer's Disease. *Am. J. Psychiatry* 141, 1356–1364. doi:10.1176/ajp.141.11.1356

Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). "A Simplest Systematics for the Organization of Turn Taking for Conversation**This Chapter Is a Variant Version of "A Simplest Systematics for the Organization of Turn-Taking for Conversation," Which Was Printed in Language, 50, 4 (1974), Pp. 696-735. An Earlier Version of This Paper Was Presented at the Conference on "Sociology of Language and Theory of Speech Acts," Held at the Centre for Interdisciplinary Research of the University of Bielefeld, Germany. We Thank Dr. Anita Pomerantz and Mr. Richard Faumann for Pointing Out to us a Number of Errors in the Text," in *Studies in the Organization of Conversational Interaction* (Elsevier), 7–55. doi:10.1016/b978-0-12-623550-0.50008-2

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53, 361–382. doi:10.2307/413107

Shriberg, E. E. (1994). "*Preliminaries To a Theory of Speech Disfluencies*,". Ph.D. thesis, Citeseer.

Siegel, S., and Castellan, N. (1988). Measures of Association and Their Tests of Significance. *Nonparametric Stat. Behav. Sci.* 13, 224–312.

Singh, S., Bucks, R. S., and Cuerden, J. M. (2001). Evaluation of an Objective Technique for Analysing Temporal Variables in Dat Spontaneous Speech. *Aphasiology* 15, 571–583. doi:10.1080/02687040143000041

Sloetjes, H., and Wittenburg, P. (2008). "Annotation by Category-Elan and Iso Dcr," in *6th International Conference on Language Resources and Evaluation* (LREC 2008).

Straiton, J. (2019). Predicting Alzheimer's Disease

Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2020). Voice Quality and Speech Fluency Distinguish Individuals with Mild Cognitive Impairment from Healthy Controls. *Plos one* 15, e0236009. doi:10.1371/journal.pone.0236009

Varela Suárez, A. (2018). The Question-Answer Adjacency Pair in Dementia Discourse. *Int. J. Appl. Linguistics* 28, 86–101. doi:10.1111/ijal.12185

Wang, C. (2019). "A Relevance-Theoretic Approach to Turn Silence," in *4th International Conference on Contemporary Education, Social Sciences and Humanities (ICCESSH 2019)* (Atlantis Press)).

Young, J. A., Lind, C., and van Steenbrugge, W. (2016). A Conversation Analytic Study of Patterns of Overlapping Talk in Conversations between Individuals with Dementia and Their Frequent Communication Partners. *Int. J. Lang. Commun. Disord.* 51, 745–756.

Zou, K. H., O'Malley, A. J., and Mauri, L. (2007). Receiver-operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation* 115, 654–657.

Conflict of Interest: The 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership