
Proceedings Paper:
Pahar, M. orcid.org/0000-0002-5926-0144, Tao, F., Mirheidari, B. et al. (8 more authors) (2025) CognoSpeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech. In: 2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM). 2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM), 17-20 Mar 2025, Trondheim, Norway. Institute of Electrical and Electronics Engineers (IEEE), pp. 1-7. ISBN: 9798331508340.

https://doi.org/10.1109/cihm64979.2025.10969487

CognoSpeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech

Madhurananda Pahar[1*], Fuxiang Tao[1*], Bahman Mirheidari[1*], Nathan Pevy[1†], Rebecca Bright[2‡], Swapnil Gadgil[2‡], Lise Sproson[3§], Dorota Braun[4*], Caitlin Illingworth[4*], Daniel Blackburn[4*], Heidi Christensen[1*]

[1]School of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK
[2]Therapy Box, London, UK
[3]NIHR Devices for Dignity HTC, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, S10 2JF, UK
[4]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, S10 2HQ, UK

Email: *{m.pahar, f.tao, b.mirheidari, d.a.braun, c.illingworth, d.blackburn, heidi.christensen}@sheffield.ac.uk, †nathanpevy@gmail.com, ‡{rbright, sgadgil}@therapy-box.co.uk, §lise.sproson@nihr.ac.uk

***Abstract*—The early signs of cognitive decline are often noticeable in conversational speech, and identifying those signs is crucial in dealing with later and more serious stages of neurodegenerative diseases. Clinical detection is costly and time-consuming and although there has been recent progress in the automatic detection of speech-based cues, those systems are trained on relatively small databases, lacking detailed metadata and demographic information. This paper presents CognoSpeak and its associated data collection efforts. CognoSpeak asks memory-probing long and short-term questions and administers standard cognitive tasks such as verbal and semantic fluency and picture description using a virtual agent on a mobile or web platform. In addition, it collects multimodal data such as audio and video along with a rich set of metadata from primary and secondary care, memory clinics and remote settings like people's homes. Here, we present results from 126 subjects whose audio was manually transcribed. Several classic classifiers, as well as large language model-based classifiers, have been investigated and evaluated across the different types of prompts. We demonstrate a high level of performance; in particular, we achieved an F_1-score of 0.873 using a DistilBERT model to discriminate people with cognitive impairment (dementia and people with mild cognitive impairment (MCI)) from healthy volunteers using the memory responses, fluency tasks and cookie theft picture description. CognoSpeak is an automatic, remote, low-cost, repeatable, non-invasive and less stressful alternative to existing clinical cognitive assessments.**

***Index Terms*—dementia, MCI, computational paralinguistics, cognitive decline, pathological speech**

I. Introduction

Struggles with memory and cognition can stem from a variety of causes, including fatigue, stress, and illness, and often worsen with age. If such issues persist beyond a few months, they might indicate mild cognitive impairment (MCI) [1], a condition linked to problems with memory, learning, reasoning, attention, conversation, language, and loss of interest or motivation. Those experiencing MCI often describe their condition as 'brain fog' as it affects their ability to think clearly [2]. Approximately half of those diagnosed with MCI eventually develop dementia, a progressive condition caused by brain-damaging diseases including Alzheimer's (AD) [3]. Other forms of dementia, such as vascular and frontotemporal dementia, often affect behaviour and cognitive abilities such as language, perceptual and executive functions [4], [5]. Both MCI and dementia are the initial stages of cognitive decline, and lack a cure [6].

The benefits of detecting early signs of dementia include timely treatment to delay the later stages. However, this can be challenging as it requires thorough investigations by neurological experts and maintaining historical records through cognitive assessment by the health service providers [7]. Existing clinical methods for diagnosis are brain magnetic resonance imaging measurement (MRI), scale testing and cerebrospinal fluid analysis [8], which are expensive, time-consuming, unpleasant to the participants and laborious; which is why these methods are disadvantageous for large-scale cognitive decline screening [9]. This is also why 75% of individuals with early cognitive decline do not get treated at all [10], making it one of the most under-diagnosed conditions for the global ageing population. There is, therefore, an increased necessity for remote, smart technologies to support healthcare services to deliver timely and accurate diagnosis [10].

CognoSpeak is designed to accelerate the identification of early cognitive decline, thus saving time and financial resources for health service providers across the world, such as the National Health Service (NHS) in the UK (Figure 1). Using a virtual agent on a laptop or tablet, CognoSpeak engages the subject in a conversation which is cognitively demanding on multiple domains involving memory, language and attention and carefully analyses the spoken answers to detect early signs of dementia as described in Figure 2. The virtual agent administers a diverse range of tasks such as long and short-term memory tests, semantic and phonemic fluency tests, picture description tasks, and reading tasks. Gold-standard clinical

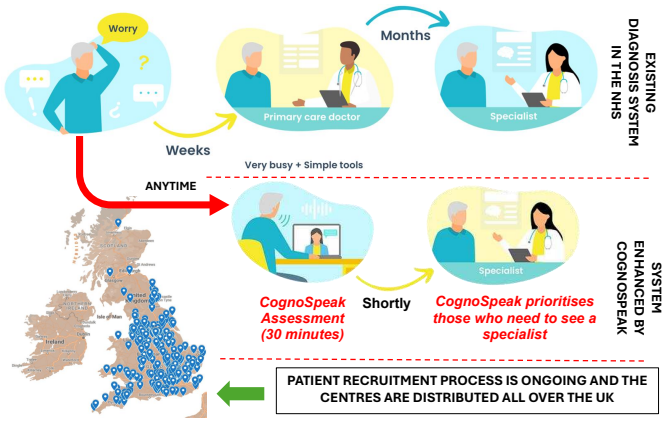

Fig. 1: **CognoSpeak** will expedite the stratification process for those who are concerned about their cognitive health as it is capable of accurately distinguishing between those who show signs of cognitive decline and therefore need referring to a more specialist assessment and those who have other causes of their memory problems such as depression or anxiety. Currently, data collection is ongoing in various parts of the UK, encompassing a wide range of accents and demographics. Participants are recruited nationwide through primary and secondary care, various websites such as Join Dementia Research and social media channels, including a number of memory clinics for the study.

processes have diagnosed the subjects' cognitive status into three categories which are dementia, MCI, and healthy control (HC). Subjects are invited for follow-up assessments to enable longitudinal studies. In addition, the subjects undergo multiple standard cognitive and mood assessments, which are further described in Section III, to allow for a range of detection and regression experimental analyses. CognoSpeak is currently used to collect data in primary and secondary care in the UK as well as recruiting nationally through websites like Join Dementia Research. This remote, large-scale data collection, either at patients' homes, at clinics or in community centres, allows us to capture real-world conversational speech. To the best of our knowledge, this is the largest collection of this type of data.

This paper presents the first results from the initial data collection phase of audio recordings from 126 subjects. Next, in Section II, similar work carried out in the past and their limitations are discussed, followed by Section III, where the system and the initially collected data are described. Section III also describes the experimental setup of feature extraction, classification and evaluation. The results are summarised in Section IV and Section V discusses them. Finally, Section VI concludes that CognoSpeak, an automatic, low-cost and remote assessment tool, is a promising viable means to provide a diagnostic aid for the early detection and tracking of signs of cognitive decline for healthcare providers. Python scripts used in this study are shared via our GitHub repository [11].

II. Previous work

Analysis of speech and conversation using machine learning techniques has shown promise in the past. High accuracies of 74.65% and 84.51% to detect AD have been achieved using both acoustic and linguistic features, respectively, from spontaneous speech [12]. In recent years, an increasing number of studies, such as those based on the DementiaBank dataset [13] and derivations of it like the ADReSS [14] and ADReSSo [15] challenges, worked on picture description data of the well-known cookie theft (CT) picture. ADReSS presents challenges of successfully identifying potential AD patients using audio recordings and manual transcriptions, whereas ADReSSo only provided researchers with the audio recordings. A special compact set of features from standalone audio helped to achieve an accuracy of 87.6% [16]. Automatic systems like [17], [18] performed very well by achieving accuracies up to around 90% using both audio and text (manual transcription) features. These public datasets have mostly supported studies on the binary distinction between HC and AD with very little work on MCI. One exception has been the work of Mirheidari *et al.* [19], which used BERT-based features combined with acoustic and textual features to train a classifier and achieve an F_1-score of 81.2% on a dataset of 50 MCI vs. 50 HC participants. Amini *et al.* [20] used AlBERT and BERT on a dataset of interviews from three groups of HC (410), MCI (387), and dementia (287) subjects to achieve an accuracy range between 62% to 69% for identifying MCI from HC. However, these datasets were small and lacked multiple diagnostic classes, as well as wider ranges of demographics including ethnicity information and rich clinical metadata.

The potential of using automatic speech analysis to monitor the progression of signs of cognitive decline is further confirmed when digital speech assessments completed by older adults over a six-month period were grouped by Montreal Cognitive Assessment (MoCA) scores and adults with higher MoCA scores showed better performance in information richness, language coherence and word-finding abilities. In contrast, MCI and AD adults demonstrated a more rapid decline [21]. As people with early cognitive decline are often not followed closely enough, automatic tools like CognoSpeak may help monitor longitudinal tracking.

III. Experimental setup

A. CognoSpeak data collection

The participants' data is collected using the CognoSpeak system, an innovative online AI tool designed to automatically identify potential indicators of cognitive decline. The assessment involves the acquisition of audio and video recordings from participants who are asked to answer a set of questions and complete a series of conventional cognitive tasks by a virtual agent. The participants can choose one of the four avatars, specially designed to represent various ethnic and age groups to make the participants feel comfortable speaking to the virtual agent (Figure 2). The virtual agent prompts have been crafted with input from clinicians and computational

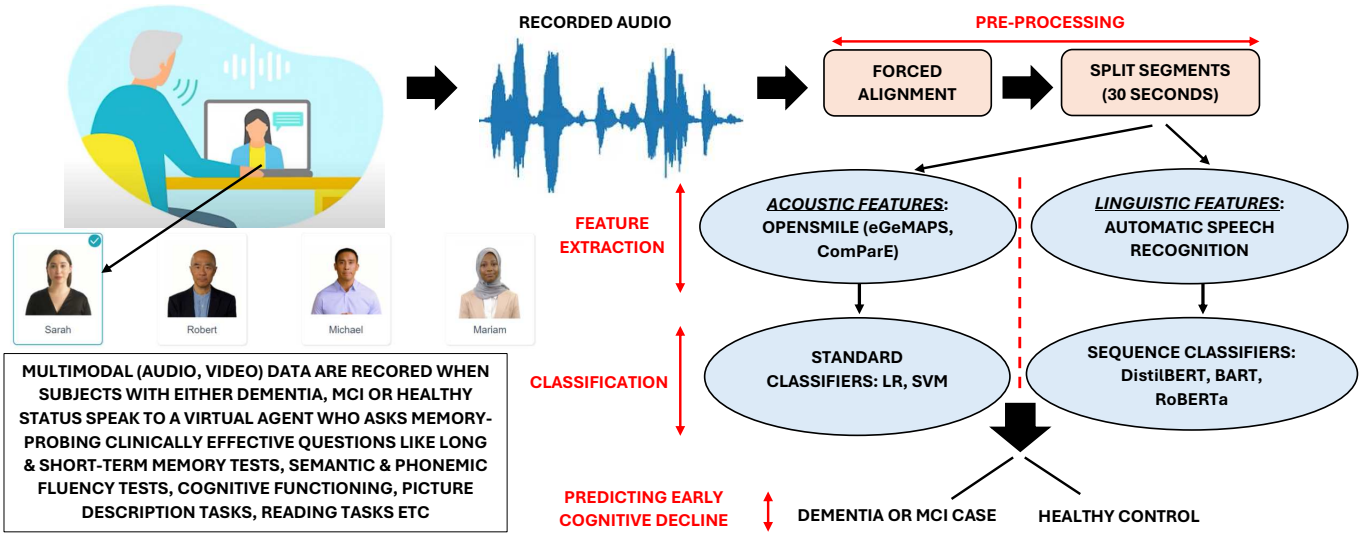

Fig. 2: **CognoSpeak** collects real-world audio and video recordings when a virtual agent prompts the subject for a diverse range of clinically proven effective tasks on multiple platforms such as mobile and web. Four avatars (2 male and 2 female) from diverse ethnic groups were used as the virtual agent. The recorded audio is then pre-processed and both acoustic and linguistic features are extracted to train and evaluate classifiers such as standard classifiers and sequence classifiers (foundation models). Finally, those with either dementia or mild cognitive impairment (MCI) are distinguished from healthy.

TABLE I: **Dataset description.** Our dataset contains 63 cases (dementia and MCI) and 63 healthy controls (HC). Acronyms *F* and *M* stand for female and male, respectively. The sum of the gender columns does not correspond to the total number of subjects (126) because 1 of these is undisclosed. The values mentioned within brackets are the standard deviations. The audio lengths are shown in seconds under 'Len' and the signal-to-noise ratios (SNR) are shown in dB. In general, audio has very little noise and the long-term prompts are shorter for dementia subjects, but are almost equal between MCI and HC.

Group	Diagnosis	Number of Subjects	Age	Gender		Short-term		Long-term		Semantic-fluency		Picture description	
				M	F	Len	SNR	Len	SNR	Len	SNR	Len	SNR
Case	Dementia	12	75.83 (9.25)	6	6	40.47 (33.79)	-84.31 (9.23)	30.53 (28.07)	-83.2 (14.4)	60.06 (0.13)	-90.56 (10.97)	83.74 (49.36)	-85.51 (7.92)
	MCI	51	68.41 (10.11)	31	20	39.12 (30.44)	-82.74 (10.88)	41.32 (28.53)	-82.33 (9.97)	60.2 (0.31)	-85.14 (10.08)	76.66 (58.96)	-88.28 (13.08)
Control	HC	63	59.19 (15.89)	28	34	41.12 (31.48)	-83.6 (15.11)	39.51 (34.73)	-83.62 (13.06)	60.18 (0.29)	-89.76 (14.76)	70.19 (35.33)	-85.17 (13.18)
Total	All	126	64.51 (14.43)	65	60	40.25 (31.3)	-83.32 (13.06)	39.39 (31.88)	-83.06 (12.06)	60.18 (0.29)	-87.96 (12.92)	74.10 (47.77)	-86.46 (12.82)

linguists to elicit diverse speech patterns and assess speech properties that may be impacted by cognitive impairments. For example, the questions include memory recall that aims to examine short or long-term memory [22], [23], speech fluency, cognitive functioning [24], picture description [25] and reading a 9-sentence or 129-word long paragraph [26].

The answers to these speech elicitation prompts are transcribed automatically to enrich the data with multiple modalities, allowing acoustic and linguistic analysis. All the interactions are also manually transcribed to aid the training of the automatic speech recognition models. In addition, some participants are required to complete a comprehensive suite of questionnaires, such as the MoCA, Multicultural Cognitive Examination (MCE), and Rowland Universal Dementia Assessment Scale (RUDAS), which are intended to evaluate cognitive impairments attributable to dementia. Mental health assessments, including the Patient Health Questionnaire (PHQ-9) and Generalised Anxiety Disorder Assessment (GAD-7), are administered to evaluate the psychological well-being of participants and to allow for the exploration of the confounding effects of cognitive decline and mood on speech parameters. In addition, to achieve cognitive diagnostic labels, participants recruited through the NHS undergo the usual gold-standard diagnosis and participants that are recruited through other routes are asked to complete the Cognitron assessment [27]. To protect privacy, participants are oriented on tool usage and briefed on each question by the virtual agent. Recruitment of participants takes place through a variety of channels, including healthcare referrals, charitable initiatives, and word-of-mouth. All subjects participate voluntarily and sign informed consent items through CognoSpeak. The data

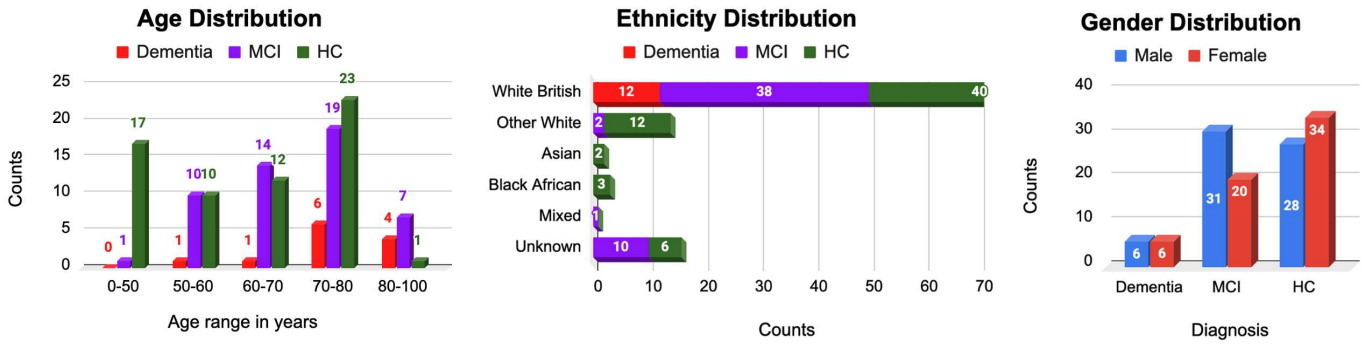

Fig. 3: **Distribution of the demographic information (age, ethnicity and gender) of all 126 subjects used in this study.** The age distribution shows the trend of having more younger people as healthy. The ethnicity distribution shows that even though a small number of subjects represent other ethnicities, most of the participants are white British. Mixed ethnicity is noted for mixed white and black. The gender distribution is almost equally balanced across the groups.

collection is performed under the ethical requirements of the institutions.

CognoSpeak has been developed with extensive user feedback throughout to ensure usefulness and usability. Users from diverse ethnic backgrounds have helped with the co-design of the virtual agents where participants have a choice of four different virtual agents. Feedback indicates that users find the system intuitive and easy to use as less than 6% of users require help to complete the assessment, and only 3.5% of users dislike using the system. In addition, nearly 96% express positive or neutral views regarding the virtual agent incorporated into the system.

B. Dataset description

Data collection is ongoing, and to date, more than 1600 participants have completed the assessment and 12.63% of them are from UK ethnic minorities for which we collect a range of language-related metadata enabling us to explore robustness to language background and potential bias issues. Of these, manual transcripts are so far available for a subset of the participants, specifically 126 individuals for the experiments conducted in this study. To explore the individual merits of the various elicitation tasks, we have conducted experiments on four sets of tasks, namely the short-term memory question, the long-term memory question, the semantic fluency task and the CT picture description task. The short and long-term memory probing questions were 'What did you do over the weekend? Please give as much detail as possible' and 'Please could you tell me about the school you went to and how old you were when you left?' The subject was asked to name as many animals as possible within a minute for the semantic fluency task and then invited to describe the CT picture. The total length of recordings for this cohort is 7 hours, 29 minutes and 13 seconds, which includes 12 individuals diagnosed with dementia, 51 diagnosed with MCI, and 63 participants identified as healthy controls, i.e., having no cognitive health issues. Table I shows the prompt lengths and signal-to-noise ratio (SNR) along with its standard deviation for each group as well. Audio in general contains very little noise as the lowest SNR has been -83.06±12.06 dB for long-term memory tasks as shown in Table I.

Table I shows the demographic information in terms of the number of subjects, age, and gender and Figure 3 shows the distribution of age, ethnicity and gender. They demonstrate that the gender distribution is almost equally balanced among both cases (dementia and MCI) and healthy controls. However, as the data has been collected mainly in England so far, the majority of the participants are white British. Figure 3 also shows that all dementia and MCI subjects are white, except 10 MCI subjects who did not disclose their ethnic information. According to a two-tailed t-test ($p < 0.001$), no statistically significant difference in gender distribution was identified, however, there is a statistically significant difference between the age distribution of patients, who were diagnosed with dementia and MCI and have an average age of 69.83±10.37 years, and healthy controls. This observation aligns with the consensus in neuroscience that cognitive decline, often attributed to structural and functional changes within the brain, predominantly occurs with aging [28].

C. Feature extraction and classification

The audio of the selected four tasks was preprocessed by forced alignment and segmentation before extracting both acoustic and linguistic features, as shown in Figure 2. Acoustic features include eGeMAPS and ComParE 2016 from OpenSmile 3.0 [29], [30]. Foundation models are proven useful in sentiment, language and text analysis in recent times [31], and here we explore their abilities in detecting cognitive impairments [32]. We have applied three such models which are BART [33], RoBERTa [34], [35] and DistilBERT [36], [37]. Initial experiments using different seeds on these models exhibited no significant improvements. The experiments were performed over k-fold cross-validation ($k = 5$) with the same data split with strictly no overlap between folds to make rigorous comparisons between approaches and make the best use of the relatively small amount of data in each class.

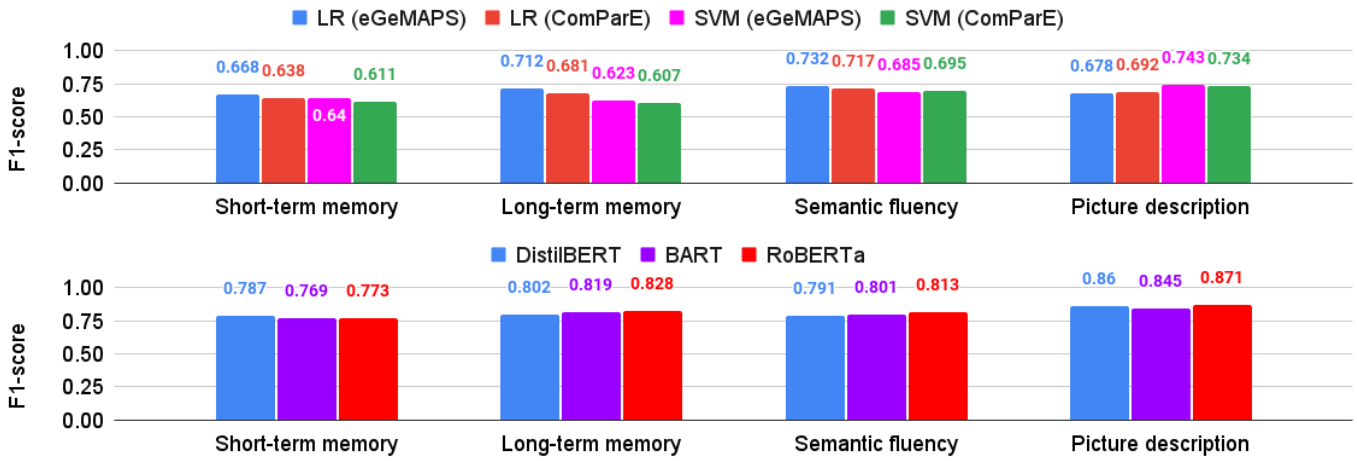

Fig. 4: **Classification based on individual task.** The F_1-scores are shown at the top and bottom half while using the acoustic and linguistic features for each classifier respectively. The performance is better when using the linguistic features.

TABLE II: **Overall classification performance.** The precision, recall and F_1-scores are shown for both acoustic and linguistic features. Again, the foundation models using linguistic features outperform the standard classifiers using acoustic features.

Experimental Setup	Features	Classifiers	Precision	Recall	F_1-score
Acoustic Features + Standard Classifiers	eGeMAPS	LR	0.682 ± 0.9	0.678 ± 0.9	0.665 ± 0.9
		SVM	0.691 ± 0.8	0.681 ± 0.8	0.684 ± 0.8
	ComParE	LR	0.691 ± 0.6	0.720 ± 0.6	0.702 ± 0.6
		SVM	0.721 ± 0.7	0.724 ± 0.7	0.723 ± 0.7
Linguistic Features + Foundation Models	ASR	DistilBERT	$\mathbf{0.875 \pm 0.2}$	$\mathbf{0.871 \pm 0.2}$	$\mathbf{0.873 \pm 0.2}$
		BART	0.834 ± 0.4	0.858 ± 0.4	0.847 ± 0.4
		RoBERTa	0.881 ± 0.3	0.864 ± 0.3	0.868 ± 0.3

IV. Results

A. Classification on individual tests

Figure 4 shows the F_1-scores achieved while using individual tasks for classification. The top half shows the results using the acoustic features while the bottom half shows the results for linguistics features.

1) Acoustic features: For the short-term memory task, a logistic regression (LR) classifier using eGeMAPS has performed the best, followed by the support vector machine (SVM), as they produced the F_1-scores of 0.668 and 0.64 respectively. Here, eGeMAPS features performed better than the ComParE features. The same pattern is also visible for long-term memory tasks as the LR using eGeMAPS again outperformed the others. It achieved an F_1-score of 0.712, and the next best performance was also achieved by LR, but this time using ComParE features. Figure 4 also shows that the variations are quite high between performances across the classifiers for using long-term memory tasks. An F_1-score of 0.732 has also been achieved from an LR for the semantic fluency task using the eGeMAPS features. ComParE features performed close to equally well by producing the F_1-score of 0.717. SVM has produced the highest F_1-score of 0.743 and 0.734 while using the eGeMAPS and ComParE features extracted from the picture description respectively. A lightly lower F_1-scores of 0.678 and 0.692 have been achieved from those features while using the LR classifier. Figure 4 demonstrates that LR has generally performed better than SVM while using acoustic features for short and long-term memory tasks, and semantic memory tasks individually.

2) Linguistic features: The bottom half of Figure 4 shows the performances of the linguistic-based foundation models. As expected, overall, these models outperform the more conventional acoustic-based classifiers. DistilBERT has performed the best while using short-term memory tasks by achieving the F_1-score of 0.787, which was closely followed by RoBERTa as it produced the F_1-score of 0.773. BART has performed similarly as its F_1-score is 0.769. However, RoBERTa has outperformed all others by achieving an F_1-score of 0.828 while using the long-term memory tasks as BART and DistilBERT have produced the F_1-scores of 0.819 and 0.802 respectively. For semantic fluency tests, RoBERTa has achieved the F_1-score of 0.813, thus outperforming others by a small margin, as BART and DistilBERT have produced the scores of 0.801 and 0.791 respectively. Finally, RoBERTa has performed the best by achieving an F_1-score of 0.871 while using the features extracted from the picture description tasks. DistilBERT and BART have also performed almost equally well by producing F_1-scores of 0.86 and 0.845.

B. Overall classification performance

The classification was also performed by considering all four tasks and applying a majority voting algorithm to predict the final label of a test subject. The precision, recall and

F_1-scores along with their standard deviations (σ) across the cross-validation folds are reported as the overall performances in Table II. The LR has produced an F_1-score of 0.665, which is slightly less than the F_1-score of 0.684 generated by the SVM using eGeMAPS features as overall performance. These performances are better while using the overall ComParE features as F_1-scores of 0.702 and 0.723 were achieved from LR and SVM. The overall performances of the foundation models using linguistic features are statistically significantly higher than classifiers using acoustic features. Here, DistilBERT have outperformed all others in overall performance by achieving the highest F_1-score of 0.873 with precision and recall values of 0.875 and 0.871 respectively with a σ of 0.2. RoBERTa has achieved an almost equal performance by generating F_1-score of 0.868 with the σ of 0.3 and BART has achieved a slightly lower F_1-score of 0.847 with the σ of 0.4.

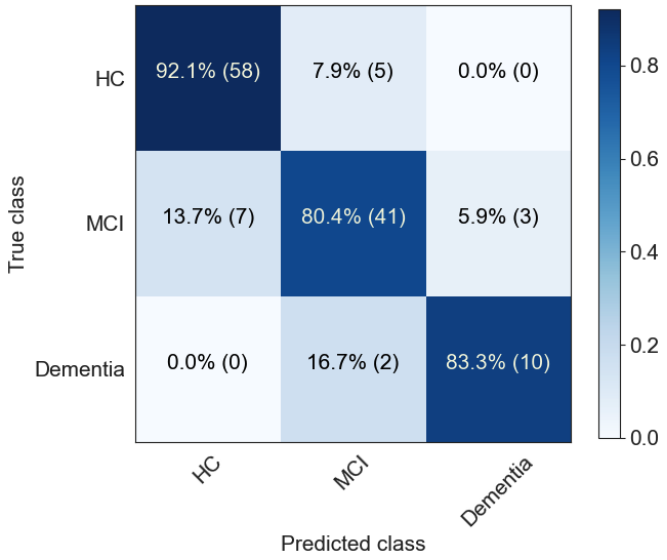

Fig. 5: **Confusion matrix.** The best-performing DistilBERT model, shown in Table II, was applied to all test folds of cross-validation. Detecting MCI was the most challenging.

V. Discussion

The results show that, in general, LR has performed better than SVM, and eGeMAPS have been the features of choice while using the acoustic features for classification. Results also provide evidence of using semantic fluency tests, long-term memory tests and picture description tasks as the primary means of detecting early signs of cognitive decline using acoustic features as they produced the highest F_1-scores of 0.743, 0.732 and 0.712 respectively. While using the overall acoustic features, ComParE has been the feature of choice as both LR and SVM performed better than eGeMAPS. However, linguistic features have outperformed the acoustic features by a large margin for both individual and overall performance. Table II also demonstrates that the σ among the cross-validation folds are lower for the linguistic models as well, indicating better generalisation over a diverse dataset. These observations also support our previous findings where a complex, pre-trained and fine-tuned architecture performed better than the standard classifiers [38].

Figure 5 presents the confusion matrix which is generated by applying the best-performed DistilBERT model, which achieved the highest F_1-score of 0.873 using all four tasks (Table II), to each cross-validation test fold and the actual and predicted labels are summed up. It exhibits good overall performance as subjects with no cognitive impairment were predicted with 92.1% accuracy, followed by 83.3% and 80.4% accuracy for MCI and dementia. This can also be translated into a two-class confusion matrix, which shows a specificity of 92.1% and sensitivity of 81% for detecting either dementia or MCI. As cognitive declines of those suffering from MCI are hard to notice, it is more challenging for the foundation models to predict accurately and hence a very promising result, despite the skewness present in the age and ethnicity distribution in our data as shown in Figure 3.

VI. Conclusion

We have presented an automatic remote assessment tool for detecting early signs of cognitive decline using real-world conversational speech. Speaking requires significant cognitive resources, including memory, language, and attention thus carrying the early signs of cognitive impairment. CognoSpeak engages participants in a conversation using a virtual agent and asks memory-probing questions alongside administering more conventional cognitive tests. To provide an initial set of results that can be used as comparisons based on the data collected by CognoSpeak in future work, a subset of 126 subjects, of whom there are 12 dementia, 51 MCI and 63 healthy controls, was prepared. Both acoustic and linguistic features were extracted from short-term memory, long-term memory, semantic fluency and picture description tasks. The linguistic features extracted from the combination of these four tasks with DistilBERT model have performed the best by producing the highest F_1-score of 0.873. The corresponding confusion matrix shows that predicting healthy controls has been the most successful, followed by dementia and MCI. CognoSpeak provides a low-cost, repeatable, non-invasive and less stressful alternative to the current cognitive assessment methods to detect early signs of cognitive decline.

As for our immediate future work, rigorous classifier training and fine-tuning will be carried out to improve the classification performance on a larger corpus, which will contain less skewed demographic information and will be extended to multi-class classification. The performance of CognoSpeak as an automatic remote long-term monitoring tool will also be investigated using the follow-up patients alongside their MoCA, MCE, RUDAS, PHQ-9 and GAD-7 scores. Finally, the entire multimodal data corpus along with the rich metadata will be released to the research community soon and a small subset of this data has already been shared with more than 80 research groups around the world as a part of the "The Prediction and Recognition Of Cognitive declinE through Spontaneous Speech (PROCESS)" signal processing grand challenge in ICASSP 2025.

REFERENCES

[1] M. Davis, T. O'Connell, S. Johnson, S. Cline, E. Merikle, F. Martenyi, and K. Simpson, "Estimating Alzheimer's disease progression rates from normal cognition through mild cognitive impairment and stages of dementia," *Current Alzheimer Research*, vol. 15, no. 8, pp. 777–788, 2018.

[2] P. B. Rosenberg, M. M. Mielke, B. S. Appleby, E. S. Oh, Y. E. Geda, and C. G. Lyketsos, "The association of neuropsychiatric symptoms in MCI with incident dementia and Alzheimer disease," *The American Journal of Geriatric Psychiatry*, vol. 21, no. 7, pp. 685–695, 2013.

[3] A. Prestia, A. Caroli, W. M. Van Der Flier, R. Ossenkoppele, B. Van Berckel, F. Barkhof, C. E. Teunissen, A. E. Wall, S. F. Carter, M. Schöll *et al.*, "Prediction of dementia in MCI patients based on core diagnostic markers for Alzheimer disease," *Neurology*, vol. 80, no. 11, pp. 1048–1056, 2013.

[4] D. S. Knopman, B. F. Boeve, and R. C. Petersen, "Essentials of the proper diagnoses of mild cognitive impairment, dementia, and major subtypes of dementia," in *Mayo Clinic Proceedings*, vol. 78, no. 10. Elsevier, 2003, pp. 1290–1308.

[5] F. Thabtah, R. Spencer, and Y. Ye, "The correlation of everyday cognition test scores and the progression of Alzheimer's disease: a data analytics study," *Health Information Science and Systems*, vol. 8, pp. 1–11, 2020.

[6] H. C. Hendrie, "Epidemiology of dementia and Alzheimer's disease," *The American Journal of Geriatric Psychiatry*, vol. 6, no. 2, pp. S3–S18, 1998.

[7] M. Shi, G. Cheung, and S. R. Shahamiri, "Speech and language processing with deep learning for dementia diagnosis: A systematic review," *Psychiatry Research*, p. 115538, 2023.

[8] Q. Yang, X. Li, X. Ding, F. Xu, and Z. Ling, "Deep learning-based speech analysis for Alzheimer's disease detection: A literature review," *Alzheimer's Research & Therapy*, vol. 14, no. 1, pp. 1–16, 2022.

[9] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux *et al.*, "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 & Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[10] S. Gauthier, P. Rosa-Neto, J. A. Morais, and C. Webster, "World Alzheimer Report 2021: Journey through the diagnosis of dementia," *Alzheimer's Disease International*, vol. 2022, p. 30, 2021.

[11] M. Pahar, "Codes for the publication "cognospeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech"," Dec. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.14515541

[12] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, "Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic-and BERT-based Alzheimer's Dementia Detection Through Spontaneous Speech." in *Interspeech*, 2021, pp. 3810–3814.

[13] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[14] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[15] ——, "Detecting cognitive decline using speech only: The ADReSS$_o$ challenge," *arXiv preprint arXiv:2104.09356*, 2021.

[16] M. R. Kumar, S. Vekkot, S. Lalitha, D. Gupta, V. J. Govindraj, K. Shaukat, Y. A. Alotaibi, and M. Zakariah, "Dementia detection from speech using machine learning and deep learning architectures," *Sensors*, vol. 22, no. 23, p. 9311, 2022.

[17] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, "Tackling the ADRESSO challenge 2021: The MUET-RMIT system for Alzheimer's Dementia Recognition from Spontaneous Speech." in *Interspeech*, 2021, pp. 3815–3819.

[18] E. Edwards, C. Dognin, B. Bollepalli, M. K. Singh, and V. Analytics, "Multiscale System for Alzheimer's Dementia Recognition Through Spontaneous Speech." in *INTERSPEECH*, 2020, pp. 2197–2201.

[19] B. Mirheidari, R. O'Malley, D. Blackburn, and H. Christensen, "Identifying people with mild cognitive impairment at risk of developing dementia using speech analysis," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–6.

[20] S. Amini, B. Hao, L. Zhang, M. Song, A. Gupta, C. Karjadi, V. B. Kolachalama, R. Au, and I. C. Paschalidis, "Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach," *Alzheimer's & Dementia*, 2022.

[21] J. Robin, M. Xu, L. D. Kaufman, and W. Simpson, "Using digital speech assessments to detect early signs of cognitive impairment," *Frontiers in digital health*, vol. 3, p. 749758, 2021.

[22] J. W. Ashford, "Screening for memory disorders, dementia and Alzheimer's disease," 2008.

[23] W. J. Lorentz, J. M. Scanlan, and S. Borson, "Brief screening tests for dementia," *The Canadian Journal of Psychiatry*, vol. 47, no. 8, pp. 723–733, 2002.

[24] G. Cipriani, S. Danti, L. Picchi, A. Nuti, and M. D. Fiorino, "Daily functioning and dementia," *Dementia & neuropsychologia*, vol. 14, pp. 93–102, 2020.

[25] K. D. Mueller, B. Hermann, J. Mecollari, and L. S. Turkstra, "Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks," *Journal of Clinical and Experimental Neuropsychology*, vol. 40, no. 9, pp. 917–939, 2018.

[26] K. Noble, G. Glosser, and M. Grossman, "Oral reading in dementia," *Brain and Language*, vol. 74, no. 1, pp. 48–69, 2000.

[27] H. Brooker, G. Williams, A. Hampshire, A. Corbett, D. Aarsland, J. Cummings, J. L. Molinuevo, A. Atri, Z. Ismail, B. Creese *et al.*, "FLAME: a computerized neuropsychological composite for trials in early dementia," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 12, no. 1, p. e12098, 2020.

[28] D. L. Murman, "The impact of age on cognition," in *Seminars in hearing*, vol. 36, no. 03. Thieme Medical Publishers, 2015, pp. 111–121.

[29] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[31] Y. Zhang, J. Gao, M. Zhou, X. Wang, Y. Qiao, S. Zhang, and D. Wang, "Text-guided foundation model adaptation for pathological image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 272–282.

[32] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[33] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[35] L. Matošević and A. Jović, "Accurate detection of dementia from speech transcripts using RoBERTa model," in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, 2022, pp. 1478–1484.

[36] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[37] A. S. Nambiar, K. Likhita, K. S. Pujya, D. Gupta, S. Vekkot, and S. Lalitha, "Comparative study of deep classifiers for Early Dementia Detection using Speech Transcripts," in *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, 2022, pp. 1–6.

[38] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features," *Computers in Biology and Medicine*, vol. 141, p. 105153, 2022. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.105153