Language and Health 2 (2024) 32–38

Contents lists available at ScienceDirect

Language and Health

journal homepage: www.sciencedirect.com/journal/language-and-health

Spoken language-based automatic cognitive assessment of stroke survivors

Bahman Mirheidari [a,*], Simon M. Bell [b], Kirsty Harkness [b], Daniel Blackburn [b], Heidi Christensen [a]

[a] University of Sheffield, Department of Computer Science,Sheffield, UK
[b] Sheffield Biomedical Research Centre, University of Sheffield and Sheffield Teaching Hospitals NHS Foundation Trust, Department of Neuroscience, Sheffield, UK

ARTICLE INFO

Keywords:
Speech Technology
Post-stroke Rehabilitation
Cognitive decline assessment

ABSTRACT

Stroke survivors (SSs) often experience cognitive decline following their initial stroke, necessitating repeat post-stroke cognitive assessments. Current methods of assessment, such as the pen-and-paper-based Montreal Cognitive Assessment (MoCA), is time-consuming and often reliant on seeing skilled clinicians in person. This is at a time when patients have a lot of often diverse rehabilitation needs. To address these challenges, our paper introduces the first system of its kind to be used for this cohort. CognoSpeak is an automated cognitive assessment system that people can use initially on the ward immediately post-stroke (baseline) and subsequently at home (follow-ups). CognoSpeak assesses cognitive decline by asking users to engage with a virtual agent by answering questions and completing clinically-motivated tasks and cognitive tests. The system then uses AI to extract and process speech, language, and interactional cues for cognitive decline. The system was originally developed for dementia; here, we show that it can successfully predict MoCA scores (regression) and identify cognitive decline predicated on a MoCA-based threshold (classification) in the stroke survivor cohort. We explore an extensive set of acoustic- and text-based features as well as different machine learning models. Leveraging a unique dataset of 55 SS CognoSpeak interactions, our findings show excellent performance for both regression and classification style prediction with the best regression result (Normalised Root Mean Squared Error (N-RMSE)) of 0.092. In addition, we show that direct classification of the MoCA score cutoff of 26 yields an F1-score of 0.74 (Specificity: 0.73, Sensitivity: 0.75) using a Logistic Regression Classifier. This demonstrates the first evidence of the system's robustness and clinical potential.

1. Introduction

Advancements in Artificial Intelligence (AI), like speech recognition technology, computer conversational interfaces and machine learning algorithms, are changing healthcare interactions. This includes diagnostic assessments, tracking of symptoms and self-management, as well as the automatic assessment of cognition using the analysis of the speech and language of people with dementia (Luz et al., 2020,Luz et al., 2021; Mirheidari et al., 2021).

We are closer than ever to a situation where intelligent conversational interfaces (embodied in robots or virtual agents) are becoming of potential use within healthcare settings (Laranjo et al., 2018; Tudor Car et al., 2020), but further research is required to understand where and how they best fit into clinical pathways. In addition, we have an improved understanding of early markers of diseases (like Alzheimer's and Parkinson's) and emotional state (e.g., low mood and anxiety) in a person's speech and language, and the associated, increasingly mature (exploiting recent advances in deep learning and large language) technologies allowing us to computerise the audio- and text-based processing and modelling. This opens up huge potential for the healthcare domain, where pathological speech processing research has been burgeoning recently.

In previous work, we developed a fully automatic system (CognoSpeak) to identify early signs of dementia through the analysis of a conversation between an Intelligent Virtual Agent (IVA) and patients with memory concerns (Mirheidari et al., 2019; 2020; 2021). The IVA prompts the users to answer a series of memory-probing questions as well as to perform some standard cognitive tests. This paper investigates the feasibility of using a variant of the CognoSpeak system to aid in the detection of possible cognitive impairment in Stroke Survivors (SSs).

Stroke is one of the leading causes of neurological disability, and it is estimated that globally, 100 million people are living with stroke. Stroke

* Corresponding author.
E-mail address: b.mirheidari@sheffield.ac.uk (B. Mirheidari).

https://doi.org/10.1016/j.laheal.2024.01.001
Received 22 November 2023; Received in revised form 15 January 2024; Accepted 27 January 2024
Available online 15 February 2024

survivors are at high risk of developing cognitive impairment, which in the vast majority of cases is likely to be classed as Vascular Cognitive Impairment (VCI). VCI leads to a sub-cortical pattern of cognitive dysfunction and is caused by recurrent stroke-like injury to the brain. Stroke can reveal latent Alzheimer's Disease (AD) (the most common cause of dementia) in the months following the stroke event Sahathevan et al. (Sahathevan et al., 2012). Post-stroke dementia is thought to account for almost half of all dementia cases (Sun et al., 2014).

In most countries, national guidelines promote early cognitive testing on all people who have had a stroke; however, current pen-and-paper-based tests, such as the Montreal Cognitive Assessment (MoCA), are not always appropriate for stroke survivors who often have motor, visual and/or language difficulties. In addition, as post-stroke cognitive impairment may not develop until several months after the acute phase of stroke, longitudinal follow-up is required to detect emerging cognitive impairment. This means that developing a method to *automatically* and accurately detect and monitor Cognitive Impairment (CI) in the early post-stroke phase can aid the management of patients (Livingston et al., 2020). In addition, the ability to do this using remote methods would greatly improve the ability for timely intervention and reduce stress and anxiety during a period when patients have experienced an acute life-altering event. Finally, it would mean that clinical resources, which are stretched in most stroke service settings (Rudd et al., 2009), could be applied to a broader range of rehabilitation needs, leading to better patient outcomes and improved satisfaction for SSs and staff.

There are multiple pen-and-paper cognitive impairment screening tests in use. One of the more widely adopted assessments is the Mini-Mental State Examination (MMSE) (Folstein et al., 1975); however, it lacks sensitivity to detect mild symptoms such as those associated with MCI Mitchell (Mitchell, 2017) and mild Vascular Cognitive Impairment (VCI). The Montreal Cognitive Assessment (MoCA) Nasreddine et al. (Nasreddine et al., 2005) is a more sensitive screening tool for the identification of MCI and VCI (Blackburn et al., 2013; Coen et al., 2016; Katz et al., 2021). MoCA scores range between 0 and 30, and the assessment includes cognitive domains such as abstraction and measuring frontal lobe function, which are more sensitive to detecting VCI.

Several studies have compared the detection rate of the MOCA to the MMSE for vascular cognitive impairment post stroke (Blackburn et al., 2013; Chen et al., 2011; Dong et al., 2010; Godefroy et al., 2011; Mai et al., 2016; Pendlebury et al., 2012;Pendlebury et al., 2010). All studies show that the MOCA is superior in both detecting vascular dementia and vascular cognitive impairment in this cohort when compared to the MMSE. It is believed that the superiority for the MOCA in detecting vascular cognitive impairment may lie in the assessment of executive functioning Mai et al. (Mai et al., 2016). Executive dysfunction can be detected in up to 77% of stroke survivors Riepe et al. (Riepe et al., 2003). Both MMSE and MOCA assess executive functioning using visuoexcutive tests of which the MOCA assessments are better at identifying subtle deficits in this area.

A growing number of studies have been aimed at developing automatic methods for predicting cognitive test scores (like MoCA and MMSE) from spoken language (Luz et al., 2021; Ostrand & Gunstad, 2021; Sun et al., 2022). However, this is an inherently sparse data domain because of ethical concerns about sharing medical recordings. As a result, most studies are based on one of the very few publicly available datasets, the DementiaBank dataset (Becker et al., 1994). It contains recordings of spontaneous speech of people with Alzheimer's Disease and Healthy Controls (HCs) describing the Cookie Theft (CT) picture and their associated MMSE scores. For instance, Yancheva et al. (Yancheva et al., 2015) used a Bayes Net regression model to predict MMSE scores based on linguistic features extracted from manual transcripts of the DementiaBank data and achieved a Mean Absolute Error (MAE) of 3.8. Fu et al. (2020) used acoustic features combined with other information, such as sex and education, to predict MMSE scores on the same dataset and reported an MAE of around 5.

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) challenge (Luz et al., 2021) proposed tasks, including MMSE prediction, on a subset of the Dementia Bank dataset. The winner of this regression task (Pappagari et al., 2021) achieved a Root Mean Squared Error (RMSE) of 3.85 on linguistic features extracted from pre-trained Bidirectional Encoder Representations from Transformers (BERT) on Automatic Speech Recognition (ASR) transcripts (Devlin et al., 2018) with the second-best system achieving an RMSE around 4.0. In Mirheidari et al. (2022), we subsequently reported results in line with the third-best team in the challenge.

The MMSE assessment is the most widely used, so only a few studies have looked at predicting MoCA scores mostly using datasets collected by individual authors. For instance, Kantithammakorn et al. (2022) aimed to automatically recognise the words uttered by 90 participants in a fluency test[1] as part of the Thai version of MoCA using an ASR that achieved a 30.8% Word Error Rate (WER). They developed an end-to-end system by combining their ASR and a Deep Neural Networks (DNN)-based model to predict their measure, the Fluency Score Accuracy (defined as the difference (ratio) between the clinicians' fluency score and the score predicted by the machine in percentage). They used the cutoff of 11 words to indicate a score of 1; otherwise, a score of 0 was assigned. After applying data augmentation and other modifications to their ASR, their best classifier achieved a 93% accuracy.

Romana et al. (2021) predicted MoCA scores for 37 people with Parkinson's disease when reading three paragraphs focusing on their speech errors in terms of disfluencies. They used the Mozilla DeepSpeech ASR (trained on LibriSpeech) and fine-tuned the ASR in a leave-one-subject-out validation approach to achieve 3.9% WER. They trained a Linear Regression (LR) model to reach an average RMSE of 2.30 on the manual transcript of the reading tasks using the error and disfluencies features. The errors increased to 2.65 when using the ASR transcripts. Compared to our work, they only focused on a three-paragraph reading task, whereas CognoSpeak includes both read and spontaneous speech. We also include a broader range of features (see Section 2.2.2), whereas this study focused mostly on detecting the number of errors that a person with cognitive decline may produce.

No work has been specifically aimed at the automatic speech-based prediction of cognitive impairment in SSs. Related work focusing on SSs has looked at ways to assess the severity of their speech impairment automatically. Liu et al. (2023) collected a database of 50 participants, 25 HCs and 25 subacute stroke patients with dysarthria due to a stroke. They extracted several audio and video features (e.g., jitter, shimmer, tongue distance, minimum/maximum internal lip distance). In their preliminary study, they found correlations between some of the features and dysarthria severity. Bandini et al. (2018) extracted a few face movement features from 12 post-stroke and 11 HC video recordings on several speech and non-speech-based tasks and trained a Support Vector Machine (SVM) classifier. The classifier achieved 87% accuracy in the orofacial impairment detection task.

Overall, there have been no studies aimed at automatically predicting MoCA scores from the speech of SS patients. In addition, many of the studies are based on manual transcripts or semi-automatic approaches. In this study, we use a fully automatic pipeline to assess SS patients' cognitive ability. We explore two different speech recognition setups: a traditional *pipeline* and a more modern large language model-based system.

The original version of CognoSpeak was targeted at patients at risk of developing dementia not related to stroke. The SS version developed for this work contains added prompts designed in collaboration with stroke medical experts.

This paper presents a version of CognoSpeak that is specifically designed for the post-stroke population. For this study, SSs have

[1] Naming as many words, beginning with a specific letter like 'k', in one minute as possible.

undertaken the CognoSpeak assessment as soon as possible after experiencing a stroke (baseline assessment). This paper uses data collected using an SS-specific version of our general CognoSpeak tool (Mirheidari et al., 2019). Participants can access the web-based tool from home or in the hospital using a laptop or tablet. Audio and video streams are recorded and saved in a remote safe server, though only the audio recordings are processed for this study. For the particular version of the system targeting the stroke population, additional prompts were introduced, including:

which helps to identify subtle low-frequency anomia to severe dysphasia by assessing both speed of speech and expressive language production in describing details of events,
which aims to assess the participant's concentration and attention.
assesses expressive language, planning and executive cognitive function.
helps to assess comprehension of language

We focus on two research questions: (i) Can we predict the MoCA scores of the participants from an analysis of their responses to the prompts initiated by the IVA (*regression*) and (ii) how accurately can we identify participants with signs of cognitive impairment (defined as having a MoCA score above a certain threshold) (*classification*)? As part of this, we present an in-depth analysis of the effect of the choice of cutoff value on classification accuracy. To the best of our knowledge, this is the first data and associated experimental research to explore the area of automatic conversation-based CI assessment for stroke survivors.

2. The post-stroke CognoSpeak system

2.1. Data

The data was collected between December 2020 and 2022 using the system either at the Royal Hallamshire Hospital, Sheffield, United Kingdom using a laptop or tablet and a headset or at participants' homes using their own devices. Ethical approval for the study was gained from the National Health Service Health Research Authority, London - Camberwell St Giles Research Ethics Committee. Approval for the study was granted on the 26th of May 2020 (REC reference: 20/LO/0376). All subjects provided written consent before entry into the study. A total of 55 SS assessments have so far been recorded. All of these assessments were undertaken as soon after the stroke event as possible. The SS participants were diagnosed with Ischaemic Stroke (36), Transient Ischaemic Attack (11) and Haemorrhagic Stroke (8), respectively. All SSs had a National Institutes of Health Stroke Scale (NIHSS) assessment and scored less than 15 (average 3.7).[2] Table 1 gives more details about the participants' demographic information.

Figure Fig. 1 shows the histogram of the MoCA scores of the participants. Scores start with a minimum of 15 and a maximum of 29 with a median of 25. The majority of the scores are between 24 and 27. A few participants completed the MoCA over the telephone (T-MoCA) during the Covid-19 pandemic.[3]

Table 1
Demographic information of the participants.

	SS (n = 55)
Age	61.1(± 15.7)
Female (%)	45.5
Age since they left education	17.1(± 3.3)
MoCA	24.3(± 3.4)
NIHSS	3.7(± 3.4)

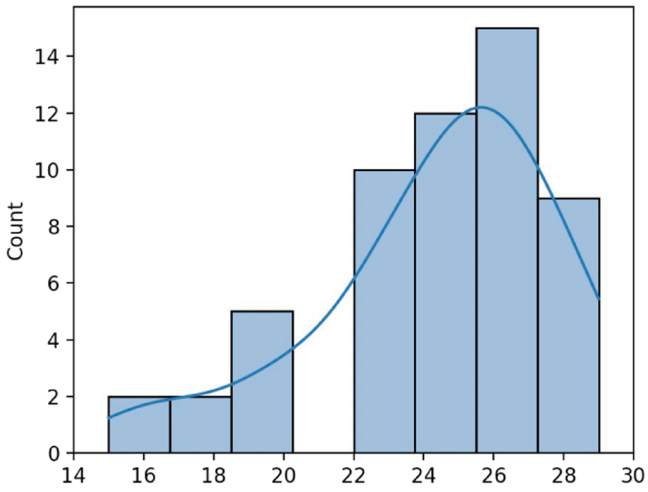

Fig. 1. Histogram of the MoCA scores of the participants. MoCA scores range between 0 and 30.

2.2. Automatic cognitive assessment system

Our automatic cognitive assessment system will predict the MoCA score based on the audio recordings of the conversations with the IVA and identify cognitive impairment. The system consists of three components: Automatic Speech Recognition (ASR), feature extraction, and a regression or classification component. Fig. 2 shows the system's structure. First, the audio files corresponding to the answers to different prompts (initiated by the IVA in the conversations) are passed to the ASR. The ASR then produces automatic transcripts, including timing information about the uttered words and the pauses between the words.[4] Next, the feature extraction component extracts different features from either the automatic transcripts or the audio files. Finally, the extracted features are passed to either a regression component to predict a MoCA score or to a classification component to identify cognitive impairment (defined as being below a certain MoCA level).

2.2.1. Automatic speech recognition

The automatic speech recognition component is trained on both publicly available data as well as in-house data. Two different models are explored: a Kaldi-based DNN model and a Wav2vec2 large language model-based framework.

For training, in addition to the CognoSpeak dataset, we used the LibriSpeech[5] dataset (over 5 thousand people reading books). Table 2 provides information about the two datasets. Note that the CognoSpeak dataset contains recordings from participants with non-SS clinical diagnoses (such as MCI and neurodegenerative disease and healthy volunteers). We used all the datasets for training the ASR system (a total of 355 recordings).

The LibriSpeech dataset was used to train a standard time delay neural network acoustic model following Kaldi's LibriSpeech recipe (Povey et al., 2011). For evaluation purposes and to mitigate the effects

[2] The NIHSS scores people on a range from 0 to 42. Scores below 5 denote either an absence of stroke symptoms or a minor stroke. Scores between 5 and 15 suggest a moderate stroke, while above 16 indicates a moderate to severe stroke.

[3] The telephone variant of MoCA is scored out of 22; we converted the T-MoCA score to the equivalent MoCA score using the guidance and conversion table provided in Katz et al. (2021)

[4] Short pauses (less than 0.1 s) are ignored.

[5] This is a large-scale state-of-the-art dataset for training high-performance ASR systems containing approximately 1000 h of read speech.

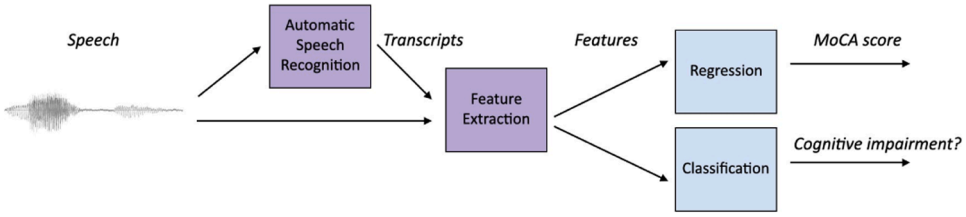

Fig. 2. Components of the automatic cognitive assessment system: automatic speech recognition (speech to transcripts), feature extraction (from speech and transcripts), regression (predicts MoCA), and classification (identifies cognitive impairment).

Table 2
Datasets used for training the ASR systems. Dur.:the total duration in hours, Avg. Dur.:average utterance duration in secs.

Dataset	Duration	# Utterances	# Speakers	Avg. Dur.
CognoSpeak	60.6 h	17.6k	355	12.4 s
LibriSpeech	961.1 h	281.2k	5466	12.3 s

of having a relatively small dataset, we used a standard k-fold cross-validation approach with 10 folds. We used the "transferring all layers" technique (both the structure and the weights were transferred) (Manohar et al., 2017), to adapt the acoustic model based on the LibriSpeech dataset to the training dataset (9 folds out of 10 of the CognoSpeak dataset). We trained the model with an audio chunk size of 100, batch size of 128, epochs of 2, a learning rate of initially 0.005 and a final learning rate of 0.0005. Two GPUs were used to train each ASR model. For the language model, a four-gram model was used with Turing smoothing interpolated with the language model based on the LibriSpeech text (50% weight for each language model). The average word error rate (WER) achieved across all participants' utterances was 25.1%. In addition, we have used the fine-tuning technique on the Wav2vec2[6] ASR (Baevski et al., 2020) to train 10 end-to-end ASRs (10-fold cross-validation). To boost the performance of the ASRs we added 5-gram language models on top of the base pre-trained Wav2vec2 model following the instructions provided by Hugging Face[7] (epochs:8, batch-size:8, learning rate:0.0001, weight decay:0.005). The final average WER was 23.2% (slightly better than Kaldi). Since the results from Wav2vec2 were not significantly better than Kaldi's we will use the outputs of both types of ASR systems in our experiments to explore any downstream effects.

While the achieved WERs are significantly higher than state-of-the-art WERs obtained on benchmark datasets and tasks, they align with similar studies in the healthcare domain (Gosztolya et al., 2019; Mirheidari et al., 2022). This speaks to the challenging nature of this data arising from the spontaneity of the speech, the age of the speakers and the nature of their condition.

2.2.2. *Feature extraction*

We have extracted a range of features from the audio recording and the outputs from the ASR systems, that is, the automatic transcripts and the word and pause timings. Table 3 summarises the different types of features. These are further described below.

Conversational analysis (CA) based features are a subset of the statistical features that we have introduced in our previous studies (Mirheidari et al., 2017, 2020) and with which we have achieved promising results when identifying cognitive impairment associated with early-stage dementia. These features are extracted across the whole conversation and were originally inspired by early work by linguists (Elsey et al., 2015) and then mapped to features that can be automatically extracted. Here, we use 10 of the originally proposed CA features, namely the features that are based on the participant's speech as opposed to that of the interviewer (here replaced with the IVA) or any accompanying person (also not available in these recordings). **Temporal features** are the pause and speech-related features extracted from the ASR transcript and the estimation of pauses between the words. For instance, the number of words, the length of words, the length of pauses, the length of the filler words, the average number of words per minute, and the average pauses per minute. The **lexical features (LX)** are 12 part-of-speech-related features from the ASR text, for instance, the number of verbs, nouns, and adverbs. The **eGeMAPS** (Eyben et al., 2015) features have successfully been used for emotion detection tasks (affective computing) (e.g., (Triantafyllopoulos et al., 2019); we have also used features proposed for the AVEC 2013 (Valstar et al., 2013) and ComPARE 2016 (Schuller et al., 2016) challenges. The **GloVe features** are the fixed-length global vectors for word representation (Pennington et al., 2014). They have been trained on a huge amount of text to capture contextual information about individual words in different text contents. In addition, we have extracted **BERT** (bidirectional encoder representations from transformers) (Devlin et al., 2018) based features that have been shown to give state-of-the-art results in many natural language processing applications, including (Pan et al., 2021). The BERT models are also trained on a huge amount of text but capture more information, such as both lexical and contextual information of words in sentences of the input text. As such, words in different order would produce different features extracted by BERT, in contrast to features based on the GloVe, which is agnostic to word order.

For the eGeMAPS, AVEC and ComPARE features, the average and standard deviation of the extracted features per frame were also calculated. For the GloVe and BERT features similarly, we calculated the average and standard deviation of the representative vectors per each word in a sentence.

Table 3
Summary of the features extracted from either the audio recording or the ASR outputs.

Features (#)	Comment
CA (10)	Inspired by Conversational Analysis (Mirheidari et al., 2017); extracted across the whole conversation, such as, number of "don't know", number of filler words.
Temporal (8)	Pause and speech-related features, e.g., the length of words, the length of pauses, the length of filler words.
LX (12)	Lexical or part of speech (e.g., number of verbs, nouns) (Mirheidari et al., 2017).
eGeMAPS (46)	Acoustic features (eGeMAPS; extended Geneva Minimal Acoustic Parameter Set) (Eyben et al., 2015) extracted using the openSMILE toolkit (Eyben et al., 2010).
AVEC13 (152)	Acoustic features (Valstar et al., 2013) extracted using the openSMILE toolkit.
ComPARE16 (260)	Acoustic features (Schuller et al., 2016) extracted using the openSMILE toolkit.
GloVe (600)	Word embedding GloVe vectors (300 dimensions)(Pennington et al., 2014).
BERT* (4096)	Bidirectional language model (large, uncased, 2048 dims) (Devlin et al., 2018).

* :BERT-Large-uncased model is available in https://huggingface.co.

[6] The Wav2vec2 model is part of a family of state-of-the-art *self-supervised* models; this one takes audio as input.

[7] https://huggingface.co/blog/wav2vec2-with-ngram

2.2.3. Regression and classification models

Since we have a relatively small amount of data to train our regression models with, we used well-known conventional machine learning regression models (with default parameters) on the extracted features to predict the MoCA scores rather than deep learning neural network models: Support Vector Regression (SVR), Logistic Regression (LR), Random Forest Regression (RFR) and K-Nearest Neighbouring Regression (KNNR). The Sklearn Python module (Pedregosa and Varoquaux, 2011) was used to train the models.

In addition, we trained binary classifiers to predict the cognitive decline at different MoCA cutoffs. Three classifiers were chosen for the classification task: logistic regression classifier (LRC), linear support vector classifier (SVC), and random forest classifier (RFC).

2.3. Evaluation

The standard metric for the regression task is the average of the Root Mean Square Error (RMSE). We used the normalised form of the metric (dividing the score by the maximum score, here 30) (N-RMSE). We applied the leave-one-subject-out cross-validation approach to evaluate the regression models' performance in our experiments (i.e., one out of 55 for the test and 54 for training, repeated for 55 recordings). Since the RFR had an internal randomness factor, we ran it ten times and calculated the average and standard deviation of the N-RMSE. Note that we have not applied fine-tuning to get better results since our data is limited; fine-tuning would involve a high risk of over-fitting.

Similarly, for classification tasks, we used the leave-one-out cross-validation approach and for evaluation, we used specificity, sensitivity and the weighted F_1-score which is again a standard machine learning performance measure for unbalanced datasets.

3. Results

Since the features were either directly extracted from the audio or the transcripts (and the word level timing information) produced by the ASRs, we first show the results on the acoustic-only features. Then we show the results on features extracted from the manual transcripts and features from the ASRs outputs.

3.1. Regression task

To answer our first research question, we have trained some regression models to estimate the MoCA scores from the extracted features. Initially, we used only-acoustic features to train the regression models.

Table Table 4 shows the average N-RMSE for the regression models on acoustic-only features. The minimum error achieved by eGeMaps features using the RFR model with an average N-RMSE of 0.095 and a standard deviation of 0.001 (we can interpret it as around a 9.5% error rate). Compare16 features were the second using the same regression model with slightly higher errors. Also, RFR worked better than other regression models using other acoustic features.

Working on the text-based features, we also observed that the RFR model outperformed other regression models. Therefore, for the rest of the regression experiments, we only show the results of this model.

Table Table 5 shows the average N-RMSE for the four regression models using the text-based features extracted from the manual transcripts, Kaldi and Wav2vec2 outputs. The BERT features achieved an average N-RMSE of 0.100 on manual transcripts. Slightly better results were achieved using Kaldi and Wav2vec2 outputs. The best performance was achieved by Wav2vec2 (0.094). We also combined all features, however, this did not result in better results. In contrast, combining BERT and eGeMaps features improved performance. The minimum N-RMSE was 0.092 achieved by Wav2vec2 on BERT + eGeMaps.

So on the whole our regression model could estimate the MoCA scores (RQ1) with a reasonably good performance.

Table 4
Average N-RMSE for the regression models using the acoustic-only features (for RFR, we list the average N-RMSE and standard deviation).

Features	SVR	LR	RFR	KNNR
AVEC13	0.116	0.169	0.104 (0.002)	0.119
ComPARE16	0.116	0.137	0.103 (0.001)	0.118
eGeMaps	0.115	0.184	**0.095** (**0.001**)	0.107

Table 5
Average N-RMSE for the RFR regression model using the text-based features extracted from the manual transcripts (Man.), Kaldi and Wav2vec2 ASRs (for RFR, Average N-RMSE (Standard Deviation)).

Features	Man.	Kaldi	Wav2vec2
CA	0.104 (0.001)	0.104 (0.001)	0.103 (0.001)
Temporal	0.108 (0.001)	0.108 (0.001)	0.106 (0.001)
LX	0.107 (0.001)	0.109 (0.001)	0.104 (0.002)
GloVe	0.104 (0.001)	0.108 (0.001)	0.104 (0.002)
BERT	0.100 (0.001)	0.096 (0.001)	0.094 (0.001)
BERT+eGeMaps	0.096 (0.002)	0.096 (0.001)	**0.092** (**0.001**)

3.2. Classification task

We further trained the three classifier models on all extracted features to predict cognitive impairment directly from the MoCA scores to explore the effect of different cutoffs varying them between 22 and 26, e.g. (MoCA $<$ 22 vs MoCA $\geq$22) (RQ2).

Table Table 6 shows sensitivity, specificity and F_1-score for the three classifiers using BERT + eGeMaps features extracted from Wav2vec2 outputs. Although the LRC and SVC achieved the maximum F_1-score of 0.86 for the MoCA cutoff 22, with a sensitivity of 0.87, their specificity is unsatisfactory (0.53). The best performance achieved was by the LRC for a cutoff score of 26 when considering the F_1-score, specificity and sensitivity altogether (F_1-score:0.74, specificity:0.73 and sensitivity:74). The same F_1-score with slightly less specificity and sensitivity was achieved by the RFC with the same cutoff.

For the LRC model with the optimum F_1-score/Specificity/Sensitivity, we calculated the confusion matrix (Fig. 3). Over 81% of MoCA scores $<$ 26 were classified correctly (25 out of 31), while only 67% of MoCA score $\geq$ 26 were classified correctly. Over 33% of those with MoCA score $\geq$ 26 with those with as MoCA score $<$ 26. Thus, in brief, the classifier models also could identify cognitive decline with an acceptable level of accuracy.

4. Discussion

In terms of predicting the MoCA score (regression task), the acoustic-only features were as important as the text-based features. In particular, the eGeMaps features achieved an N-RMSE of 0.095 on the RFR model. These features are designed to capture emotion in speech efficiently. In terms of text-based features, the BERT features were more important. However, the combination of BERT and eGeMaps achieved the best performance with an N-RMSE of 0.092. The BERT features capture linguistics and contextual information of the text. Thus, both emotion-based acoustic features, as well as linguistic features are shown to help in predicting the MoCA scores from the conversations. The prompt selection investigation, however, showed that similar results could be achieved only using five prompts.

A MoCA cutoff score of less than 26 is appropriate for detecting MCI. However, no defined normality cutoff score has been validated in the post-stroke population. Godefroy et al. (2011) used a cutoff of 26 and found that 82% of patients with acute stroke scored below this number. However, when they adjusted the scores based on age and education, the figure dropped to 48%, reaching a high sensitivity (94%) but a low specificity (42%). An optimal cutoff of 19 for non-adjusted MoCA and 22

Table 6
Classification results in terms of F_1-score (Specificity/Sensitivity) for the three models: RFC, LRC and SVC with different MoCA cutoffs.

Cuttoff	RFC			LRC			SVC		
22	0.84	0.43	0.86	**0.86**	0.53	0.87	**0.86**	0.53	0.87
23	0.75	0.40	0.78	0.73	0.39	0.75	0.72	0.44	0.73
24	0.68	0.55	0.70	0.73	0.72	0.73	0.68	0.59	0.69
25	0.66	0.60	0.67	0.65	0.62	0.66	0.69	0.66	0.69
26	0.74	0.72	0.74	**0.74**	**0.73**	**0.75**	0.71	0.70	0.71

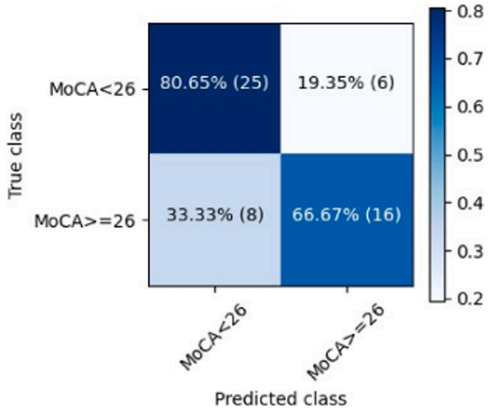

Fig. 3. Confusion matrix of the LRC model with the optimum F_1-score/Specificity/sensitivity on Table 6.

for age and education was suggested when the MoCA score was adjusted. Other proposed cutoffs included (Salvadori et al., 2013) reporting a sensitivity of 91% and specificity of 75% using a cutoff 21. Our study shows how considering different cutoff values might result in better specificity whilst losing sensitivity. However, the results still justify an optimum cutoff of 26, given both sensitivity and specificity are above 0.73.

This study has highlighted some important findings and implications for how future studies might be conducted. Most of the SS participants found the full assessment session lengthy, which may have been a contributing factor to participants either tiring through the session or deciding to not sign up to the study. We suspect they may have preferred to have a conversation with fewer questions. These limitations suggest making a shorter version of the system specifically designed for the SS. It should also be considered that the timing of the assessment may have affected study participation. When the SS clinical situation is more stable and the patients are less worried about their immediate health situation they may be more amenable to the assessment.

5. Conclusions

To the best of our knowledge, this study is one of the first to automatically predict the full MoCA cognitive test scores from the audio recordings of conversational speech of stroke survivors. The automatic features extracted from the audio recordings and the ASR outputs were largely useful in predicting the scores using the regression models. We have tried four different regression models. However, the best experiment was the Random Forest Regression model, which achieved an average N-RMSE of 0.092 on BERT from Wav2vec2 outputs plus eGeMaps features. This confirms that despite the errors introduced by the ASR, the language model-based BERT features are robust and able to capture information, which is useful in predicting the MoCA score.

Overall, we achieved very encouraging results for this study. However, a delimitating factor was the relatively small number of samples, which precluded us from using state-of-the-art DNN-based models. The SSs included in this study were predominantly mild stroke, which is largely representative of the post-stroke cohort, and they represent a group that has a considerable risk of developing cognitive impairment (Pendlebury et al., 2011). We continue to collect more data (including follow-up assessments), allowing us to explore more complex models to detect cognitive impairment and monitor decline over time.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Bahman Mirheidari reports financial support was provided by The Rosetrees Trust and the Stoneygate Trust (COMPASS, Grant Agreement No. M934), and NIHR Academic Clinical Lectureship in Neurology (CL-2020–04-00). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

This work is supported by the Rosetrees Trust and the Stoneygate Trust (COMPASS, Grant Agreement No. M934). An NIHR Academic Clinical Lectureship in Neurology CL-2020–04-004 NIHR supports SMB. This summarises independent research at the NIHR Sheffield Biomedical Research Centre (Translational Neuroscience).

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460.

Bandini , A., Green , J.R., Richburg , B., Yunusova , Y., (2018). Automatic detection of orofacial impairment in stroke.In: Interspeech, pp. 1711–1715.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology, 51*(6), 585–594.

Blackburn, D. J., Bafadhel, L., Randall, M., & Harkness, K. A. (2013). Cognitive screening in the acute stroke setting. *Age and Ageing, 42*(1), 113–116.

Chen, C., Dong, Y., Venketasubramanian, N., Sharma, V., Chan, B., Teoh, H., Seet, R., Slavin, M., Sachdev, P., Collinson, S., et al. (2011). A comparison of the baseline montreal cognitive assessment (moca) and the baseline mini-mental state examination (mmse) in predicting moderate to severe poststroke cognitive impairment. *Cerebrovascular Diseases, 32*, 44–45.

Coen, R. F., Robertson, D. A., Kenny, R. A., & King-Kallimanis, B. L. (2016). Strengths and limitations of the MoCA for assessing cognitive functioning: Findings from a large representative sample of irish older adults. *Journal of Geriatric Psychiatry and Neurology, 29*(1), 18–24.

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

Dong, Y., Sharma, V. K., Chan, B. P.-L., Venketasubramanian, N., Teoh, H. L., Seet, R. C. S., Tanicala, S., Chan, Y. H., & Chen, C. (2010). The montreal cognitive assessment (moca) is superior to the mini-mental state examination (mmse) for the detection of vascular cognitive impairment after acute stroke. *Journal of the Neurological Sciences, 299*(1-2), 15–18.

Elsey, C., Drew, P., Jones, D., Blackburn, D., Wakefield, S., Harkness, K., Venneri, A., & Reuber, M. (2015). Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient Education and Counseling, 98*, 1071–1077.

Eyben , F., Wöllmer , M., Schuller , B., (2010). Opensmile: The Munich versatile and fast open-source audio feature extractor, In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), 190–202.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189–198.

Fu , Z., Haider , F., Luz , S., (2020). Predicting mini-mental status examination scores through paralinguistic acoustic features of spontaneous speech, In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).IEEE, 5548–5552.

Godefroy, O., Fickl, A., Roussel, M., Auribault, C., Bugnicourt, J. M., Lamy, C., Canaple, S., & Petitnicolas, G. (2011). Is the montreal cognitive assessment superior to the mini-mental state examination to detect poststroke cognitive impairment? A study with neuropsychological evaluation. *Stroke, 42*(6), 1712–1716.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & Hoffmann, I. (2019). Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features. *Computer Speech & Language, 53*, 181–197.

Kantithammakorn, P., Punyabukkana, P., Pratanwanich, P. N., Hemrungrojn, S., Chunharas, C., & Wanvarie, D. (2022). Using automatic speech recognition to assess Thai speech language fluency in the Montreal cognitive assessment (MoCA). *Sensors, 22*(4), 1583.

Katz, M. J., Wang, C., Nester, C. O., Derby, C. A., Zimmerman, M. E., Lipton, R. B., Sliwinski, M. J., & Rabin, L. A. (2021). T-moca: A valid phone screen for cognitive impairment in diverse community samples. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 13*(1), Article e12144.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y., et al. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association, 25*(9), 1248–1258.

Liu, J., Du, X., Lu, S., Zhang, Y.-M., An-ming, H., Ng, M. L., Su, R., Wang, L., & Yan, N. (2023). Audio-video database from subacute stroke patients for dysarthric speech intelligence assessment and preliminary analysis. *Biomedical Signal Processing and Control, 79*, Article 104161.

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., et al. (2020). Dementia prevention, intervention, and care: 2020 report of the lancet commission. *The Lancet, 396*(10248), 413–446.

Luz , S., Haider , F., De la Fuente , S., Fromm , D., MacWhinney , B., (2020). Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge, arXiv preprint arXiv:2004.06833, 2020.

Luz , S., Haider , F., De La Fuente , S., Fromm , D., MacWhinney , B., (2021). Detecting cognitive decline using speech only: The ADReSSo challenge, arXiv preprint arXiv: 2104.09356, 2021.

Mai, L. M., Sposato, L. A., Rothwell, P. M., Hachinski, V., & Pendlebury, S. T. (2016). A comparison between the moca and the mmse visuoexecutive sub-tests in detecting abnormalities in tia/stroke patients. *International Journal of Stroke, 11*(4), 420–424.

Manohar , V., Povey , D., Khudanpur , S., (2017). JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning, In: Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, pp. 346–352.

Mirheidari , B., Blackburn , D., Christensen , H., (2022). Automatic cognitive assessment: Combining sparse datasets with disparate cognitive scores, in Proc. Interspeech. ISCA, 2022.

Mirheidari, B., Pan, Y., Blackburn, D., O'Malley, R., Christensen, H., (2021). Identifying cognitive impairment using sentence representation vectors, Proc. Interspeech, pp. 2941–2945.

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., & Christensen, H. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, 1–15.

Mirheidari , B., Blackburn , D., O'Malley , R., Walker , T., Venneri , A., Reuber , M., Christensen , H., (2019). Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia, In: Proceedings of the ICASSP. IEEE, pp. 2732–2736.

Mirheidari, B., Blackburn, D., O'Malley, R., Venneri, A., Walker, T., Reuber, M., & Christensen, H. (2020). Improving cognitive impairment classification by generative neural network-based feature augmentation. *Proc. Interspeech*, 2527–2531.

Mitchell, A. J. (2017). The mini-mental state examination (MMSE): update on its diagnostic accuracy and clinical utility for cognitive disorders. *Cognitive screening instruments* (pp. 37–48). Springer,.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*(4), 695–699.

Ostrand, R., & Gunstad, J. (2021). Using automatic assessment of speech production to predict current and future cognitive function in older adults. *Journal of Geriatric Psychiatry and Neurology, 34*(5), 357–369.

Pan , Y., Mirheidari , B., Harris , J.M., Thompson , J.C., Jones , M., Snowden , J.S., Blackburn , D., Christensen , H., (2021). Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech, Proc. Interspeech, pp. 3810–3814.

Pappagari, R., Cho, J., Joshi, S., Moro-Velazquez, L., Zelasko, P., Villalba, J., & Dehak, N. (2021). Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios. *Proc. Interspeech*.

Pedregosa, F., & Varoquaux, G. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pendlebury, S. T., Cuthbertson, F. C., Welch, S. J., Mehta, Z., & Rothwell, P. M. (2010). Underestimation of cognitive impairment by mini-mental state examination versus the montreal cognitive assessment in patients with transient ischemic attack and stroke: A population-based study. *Stroke, 41*(6), 1290–1293.

Pendlebury, S. T., Wadling, S., Silver, L. E., Mehta, Z., & Rothwell, P. M. (2011). Transient cognitive impairment in tia and minor stroke. *Stroke, 42*(11), 3116–3121.

Pendlebury, S. T., Mariz, J., Bull, L., Mehta, Z., & Rothwell, P. M. (2012). Moca, ace-r and mmse versus the ninds-csn vci harmonisation standards neuropsychological battery after tia and stroke. *Stroke, 43*(2), 464.

Pennington , J., Socher , R., Manning , C., (2014). Glove: Global vectors for word representation, In: Proc. EMNLP, pp. 1532–1543.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, The Kaldi speech recognition toolkit, In: Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.

Riepe, M. W., Riss, S., Bittner, D., & Huber, R. (2003). Screening for cognitive impairment in patients with acute stroke. *Dementia and Geriatric Cognitive Disorders, 17*(1-2), 49–53.

Romana , A., Bandon , J., Perez , M., Gutierrez , S., Richter , R., Roberts , A., Provost , E. M., (2021). Automatically detecting errors and disfluencies in read speech to predict cognitive impairment in people with Parkinson's Disease, In: Proceedings of the INTERSPEECH 2021. International Speech Communication Association, pp. 156–160.

Rudd, A. G., Jenkinson, D., Grant, R. L., & Hoffman, A. (2009). Staffing levels and patient dependence in english stroke units. *Clinical Medicine, 9*(2), 110.

Sahathevan, R., Brodtmann, A., & Donnan, G. A. (2012). Dementia, stroke, and vascular risk factors; A review. *International Journal of Stroke, 7*(1), 61–73.

Salvadori, E., Pasi, M., Poggesi, A., Chiti, G., Inzitari, D., & Pantoni, L. (2013). Predictive value of moca in the acute phase of stroke on the diagnosis of mid-term cognitive impairment. *Journal of Neurology, 260*(9), 2220–2227.

Schuller , B., Steidl , S., Batliner , A., Hirschberg , J., Burgoon , J.K., Baird , A., Elkins , A., Zhang , Y., Coutinho , E., Evanini , K., (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language, In: Proceedings of the 17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1–5, vol. 8. ISCA, vol. 8. ISCA, 2001–2005.

Sun, J.-H., Tan, L., & Yu, J.-T. (2014). Post-stroke cognitive impairment: epidemiology, mechanisms and management. *Annals of Translational Medicine, 2*(8).

Sun , L., Zheng , J., Li , J., Qian , C., (2022). Exploring mmse score prediction model based on spontaneous speech.In: SEKE, 347–350.

Triantafyllopoulos , A., Keren , G., Wagner , J., Steiner , I., Schuller , B., (2019). Towards robust speech emotion recognition using deep residual networks for speech enhancement.

Tudor Car, L., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, Y.-L., & Atun, R. (2020). Conversational agents in health care: Scoping review and conceptual analysis. *Journal of Medical Internet Research, 22*(8), Article e17158.

Valstar , M., Schuller , B., Smith , K., Eyben , F., Jiang , B., Bilakhia , S., Schnieder , S., Cowie , R., Pantic , M., (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, 3–10.

Yancheva , M., Fraser , K.C., 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, 134–139.