



# Automatic cognitive assessment: Combining sparse datasets with disparate cognitive scores

Bahman Mirheidari<sup>1</sup>, Daniel Blackburn<sup>2</sup>, Heidi Christensen<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Sheffield, UK

<sup>2</sup>Sheffield Institute for Translational Neuroscience (SITraN); Department of Neuroscience, University of Sheffield, UK

b.mirheidari@sheffield.ac.uk, heidi.christensen@sheffield.ac.uk

## Abstract

Automatic prediction of cognitive assessment scores through analysis of speech is a challenging task not least due to the lack of available data; this is exacerbated by datasets often being accompanied by disparate cognitive scores as diagnostic practices vary across the world. The ADReSSo 2021 challenge aimed at supporting research in this area and defined a number of tasks including a regression task (predicting Mini-Mental State Examination (MMSE) scores). It saw the successful introduction of a number of BERT-based models including our winning classification approach that successfully applied data augmentation using ASR-generated hypotheses. In this paper, we port this approach to the regression task and further present an investigation into the effect of combining smaller datasets with disparate cognitive scores. In particular, we combine the ADReSSo data with our in-house IVA dataset, which is associated with a different type of cognitive assessment: the Addenbrooke's Cognitive Examination (ACE-III). We show improved performance by converting ACE-III to MMSE scores thus enabling us to combine the two datasets. By selecting good hyper-parameters, the RMSE reduces from 4.45 to **4.40** on the ADReSSo task. Likewise, using the ADReSSo dataset to boost the IVA regression model, decreases RMSE from 3.50 to **3.00**.

**Index Terms:** clinical applications of speech technology, sparse data, automatic speech recognition, data augmentation

## 1. Introduction

Dementia affects individuals' cognitive skills, memory, language, speech and communication. The number of people developing dementia is increasing drastically. At the moment, there are around 850,000 people living with dementia in the UK and it is predicted that this figure will rise in the future [1]. Early detection of dementia is a challenging task due to overlapping symptoms with normal ageing, and the limited capability of existing screening tools. Speech and language abilities are routinely assessed in current cognitive tests, and studies have shown promise for the automatic extraction of cues from speech for detecting cognitive decline (e.g. [2, 3, 4]). In particular, we have developed a fully automatic system to identify dementia through the analysis of conversation between an Intelligent Virtual Agent (IVA) and patients [5, 6, 4]. The IVA prompts the users to answer a number of questions as well as to perform some standard cognitive tests including the Cookie Theft (CT) picture description task [7, 8].

There is growing interest in exploring ways of predicting some of the standard cognitive test scores directly from the speech data. However, data limitations and ethical issues with sharing medical data hinder this work. Another issue is that

different datasets may be released with different cognitive assessment scores making pooling of the data for modelling challenges. Most of the studies have been based on the Dementia-Bank dataset [9]. That dataset contains Cookie Theft (CT) picture descriptions and associated Mini-Mental State Examination (MMSE) scores (one of the standard cognitive assessment scores; more detail in Section 2). For instance, [10] worked on predicting the MMSE scores from the manual transcripts using linguistic features on DementiaBank data, and [11] used acoustic features combined with other information such as sex and education to predict the MMSE scores.

Recently, the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) challenge was organised. The main purpose of the challenge was to foster research in the automatic detection of Alzheimer's Dementia (AD) [3]. The challenge provided audio recordings of people describing the CT picture (no manual transcriptions) to perform three different tasks: i) detecting ADs from non-ADs, ii) estimating the MMSE scores, iii) identifying the progress of dementia from the recordings of previous sessions.

One of our five proposed models for the challenge [12] was among the three joint winners of task one (the two others were [13, 14]) achieving an 84.51% accuracy rate. In this paper, we aim to build a regression model for task two based on our successful task one model using the Bidirectional Encoder Representations from Transformers (BERT) [15] on different hypotheses produced by an Automatic Speech Recognition (ASR) system. However, from the results of the published papers in the challenge, we can observe two issues: firstly, the performances of the models on the evaluation set were not always replicated on the test set (compare the results on the evaluation and test sets [14, 12]), secondly, the best classification models did not gain the best results when used for regression (e.g. the Global fusion model of [13] with a classification accuracy of 84.51% (joint winner) only achieved a Root Mean Square Error (RMSE) of 4.62, which was less than their best regression model with RMSE of 3.85).

We know that the data augmentation techniques can generally improve the performance of Deep Neural Network (DNN)-based regression models. However, the participants in our in-house IVA dataset have not had MMSE assessments but have instead undergone the Addenbrooke's Cognitive Examination (ACE), which is regarded as the superior cognitive assessment task in the UK. This cognitive test (administered normally by clinical experts) is different to the MMSE, including having two different scales (range of values).

So in this study, we are going to investigate two research questions: 1) Could our BERT-based model, introduced in the ADReSSo 2021 challenge task one, be expanded to a corre-

sponding regression model for task two? 2) Is it possible to mix data with different cognitive test scales (MMSE and ACE-III), and how will that affect the performance of the regression model? To the best of our knowledge, this is the first work exploring regression modelling combining two datasets with disparate cognitive tests for regression modelling.

The rest of the paper is structured as follows. Section 2 provides a brief background of the related work. Section 3 includes the experimental setup, covering the details of the ASR and the regression models. Sections 4 and 5 contain the results and the conclusions.

## 2. Cognitive assessment tests

As mentioned above, there are a number of widely used cognitive assessment tests. The Mini-Mental State Examination (MMSE) [16] test focuses on six categories of abilities including orientation, registration, attention/calculation, recall, language and coping. The scores of each category are between zero and five (maximum score =30). People with a score  $\leq 17$  are considered as having a severe cognitive impairment, scores between 18 and 23 indicate Mild Cognitive Impairment (MCI), and scores  $\geq 24$  are taken to mean no cognitive impairment. The MMSE test is simple to administer and therefore very popular as part of a general diagnostic procedure or on its own. [17]). However, some authors are concerned about its validity and specificity [18]. Scores from this test are often used as *gold standard* labels for research datasets in this area including the publicly available DementiaBank and ADReSSo datasets.

In the UK, a widely used test is the Addenbrooke’s Cognitive Examination (ACE) [19]. It has 100 points focusing on five cognitive skills: attention/orientation, memory, language, verbal fluency, and visuo-spatial ability. It takes more time to undergo than the MMSE and requires a higher level of familiarity with cognitive disorders to administer and score. The ACE-R [20] is a revised version of the ACE with more clear domain scores, and the most recent version of the ACE is ACE-III which adds some similar items from the MMSE to the ACE assessment [21]. Our in-house IVA dataset has ACE-III scores associated with each recording.

To ease comparison between the different cognitive tests, conversion scales have been proposed with [22] providing tables enabling conversions between ACE-III and MMSE scores. We will use the same conversion approach (between ACE-III and MMSE) on the IVA dataset.

## 3. Experimental setup

We will use two main datasets to explore to what degree data collected with different cognitive assessment scores can be merged to improve score prediction: the IVA dataset (with ACE-III scores) and the ADReSSo dataset (with MMSE scores).

The IVA data was collected between 2016 and 2020 at the Department of Neurology, University of Sheffield, the UK based at the Royal Hallamshire Hospital. Out of 168, 121 participants had ACE-III test scores (59 healthy controls (HC), 20 with neurodegenerative dementia (ND), 18 with MCI, and 24 others). The ADReSSo data contain 166 training audio records with corresponding MMSE scores, as well as 71 test records. Most of the records are from HC participants and people with Alzheimer’s (these details were not shared by the organisers of the challenge).

### 3.1. Automatic speech recognition

Most approaches for automating either dementia detection or cognitive score prediction rely on both acoustic- and text-based features, which for the latter means using an ASR system to transcribe the spoken language into text.

For the ADReSSo challenge dataset, we already trained an ASR using different datasets (e.g. LIBRISPEECH, AMI - for more details refer to [12]). However, as no manual transcripts were released with the ADReSSo challenge, we could not measure the real Word Error Rate (WER).

For the IVA dataset, we used 168 audio recordings and 10-fold cross-validation using the transfer learning technique on the LIBRISPEECH dataset following Kaldi’s [23] TDNN recipe. Table 1 provides more details about the datasets used for training the ASRs. The average WER achieved was 27.89%. For training the language models, we used the interpolation approach between the text of the two datasets with four-grams and Turing smoothing.

Table 1: *Datasets used for training the ASR. Len.:the total length in hours/mins, Utts.:number of utterances, Spks.:number of speakers, and Avg.Utts.:Average utterance length in seconds.*

Dataset (No)	Len.	Utts.	Spks.	Avg. Utts.
IVA (168)	26.7h	8.3k	219	11.5s
LIBRISPEECH (281241)	961.1h	281.2k	5466	12.3s

### 3.2. Regression model

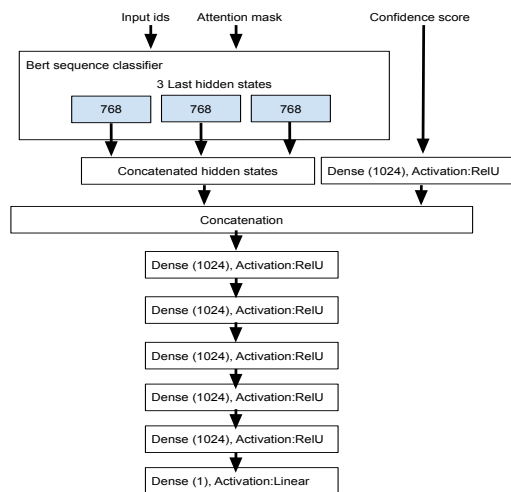


Figure 1: *Structure of the BERT-based regression model; see text for details.*

BERT models are employed in many state-of-the-art Natural Language Processing (NLP) applications and have recently been used for dementia detection [24, 4]. There has been a number of studies using BERT in the ADReSS challenge 2020 [2], in which the manual transcripts of the recording were available. For instance, [25] reported an RMSE of 4.63 using DISTIL-BERT and [26] achieved RMSE of 5.32 using BERT on transcripts combined with the information of silence and acoustic features.

The best three performances in the ADReSSo challenge 2021 for task two were [13, 14, 27] with an RMSE of 3.85, 4.35, and 4.44 respectively, the baseline RMSE achieved by

the organisers’ regression system was 5.28. To build our regression model we used a similar approach as we used for the ADReSSo task one [12] (Model 4). A range of ASR hypotheses was derived from the decoder lattices using different language weights and word insertion penalties (weights between 6 and 10 with word insertion penalties of 0, 0.5 and 1). So 30 different hypotheses were produced for each recording. The average confidence scores of the words were calculated for each of the hypotheses. Using a variety of outputs from the ASR alongside the corresponding confidence scores helped the network to be trained more robustly on the words produced by the ASR.

From the ‘uncased BERT’ sequence classifier, the three last layers were concatenated and then mixed with the input confidence scores, followed by five dense layers and a final single dense with linear activation function (the other activation functions were ReLU). Figure 1 shows the structure of one model, in which the number of neurons in the dense layers (dims) are 1024. For consistency, we use a similar structure for all the regression models in this study, except we change the number of dims and the maximum length of words passed to the BERT classifier (max-length). Also to train the models we used three epochs and a batch size of 8 with 5% of the training data used as the evaluation set. We observed that having more epochs caused the over-fitting issue. The best model was chosen from the three results on the evaluation set in the epochs (i.e., the one with the minimum RMSE). Note that due to GPU limitations we used the more lightweight ‘bert-base-uncased’, and we also did not add dropout layers in the model as this causes non-deterministic results and the need to repeat training (e.g. 10 times) and average the results.

## 4. Results

In the following section, we present the results gained from running the regression models on the ADReSSo and the IVA datasets individually. These will form the baseline for the experiments where the data is combined and trained with converted cognitive assessment scores. To measure the performance of the regression models the standard RMSE was used (equation 1). In addition, for comparison between two cognitive tests with different scales, we use the normalised RMSE (N-RMSE) (dividing RMSE by the range between the minimum score ( $y_{min}$ ) and maximum score ( $y_{max}$ )) (Note that there are other N-RMSE such as dividing by the mean, standard deviation; for its simplicity and intuition we used equation 2).

$$RMSE = \sqrt{\frac{\sum_1^n (y - \hat{y})^2}{n}} \quad (1)$$

$$N - RMSE = \frac{RMSE}{y_{max} - y_{min}} \quad (2)$$

where  $y$  is the real label,  $\hat{y}$  is the estimated label, and  $n$  is the number of samples.

### 4.1. Regression on the ADReSSo dataset

We passed the transcript hypotheses and the confidence scores as output by the ASR system for the ADReSSo data to the regression model resulting in predicted MMSE scores. Since there were multiple hypotheses per recording, we averaged the MMSE scores to get the final MMSE. Table 2 shows RMSE and N-RMSE on regression models trained using three different values for dims (1024, 512, 256) and three maximum word lengths (110, 105, 100). We chose these ranges of values due to time,

Table 2: ADReSSo RMSE of the regression models predicting MMSE (maximum score:30). max-len: maximum length of words.

Row	dims	max-len	RMSE	N-RMSE
1	1024	110	4.56	0.1520
2	1024	105	4.73	0.1577
3	1024	100	4.78	0.1593
4	512	110	4.53	0.1510
5	512	105	4.60	0.1533
6	512	100	4.62	0.1540
7	256	110	4.71	0.1570
8	256	105	<b>4.45</b>	<b>0.1483</b>
9	256	100	4.68	0.1560

Table 3: IVA RMSE of the regression models predicting ACE-III (maximum score:100). max-len: maximum length of words.

Row	dims	max-len	RMSE	N-RMSE
1	1024	110	12.90	0.1290
2	1024	105	<b>12.81</b>	<b>0.1281</b>
3	1024	100	12.93	0.1293
4	512	110	13.22	0.1322
5	512	105	13.02	0.1302
6	512	100	13.39	0.1339
7	256	110	13.21	0.1321
8	256	105	13.02	0.1302
9	256	100	13.25	0.1325

resource limitations and for making better comparisons between the models that we built. The higher these hyper-parameters values, the more required memory and computation resources. In addition, having a longer word length or more dims do not necessarily improve the performance of the regression model. As can be seen, the model with 256 dims and the maximum word length of 105 achieved an MMSE of 4.45 (comparable to the third-ranked system in the ADReSSo -2021 regression task [27] with RMSE of 4.44). The model achieved an N-RMSE of 0.1483.

### 4.2. Regression on the IVA dataset

For the IVA dataset, we used 2-fold cross-validation to divide the data into train and test sets and then measure the scores in two folds. So, we passed the transcript hypotheses and the confidence scores of the ASR to the regression model to predict the ACE-III scores. Table 3 shows the RMSE and N-RMSE results for this regression. The best model had 1024 dims and a maximum word length of 105 with RMSE of 12.81 (N-RMSE of 0.1281 or 12.81% error). Comparing the N-RMSE scores with Table 2, the performances of the regression models seem better. This reflects that the IVA dataset was less challenging than the ADReSSo in terms of the audio quality and also due to the availability of the human manual transcripts that allowed to train more robust ASRs.

Since the scale of ACE-III is different from MMSE, we used the conversion technique of [22] to estimate the equivalent MMSE scores. This approach has been shown to have high reliability (between 90% to 94%). Then we repeated the regression task with the converted scores and gained the results listed in Table 4. The conversion resulted in better N-RMSE scores and the best model with 512 neurons and the maximum word length of 110 achieved 0.1143 N-RMSE and 3.43 RMSE.

Table 4: RMSE of the regression models predicting converted MMSE (maximum score:30) on the IVA dataset. max-len: maximum length of words.

Row	dims	max-len	RMSE	N-RMSE
1	1024	110	3.53	0.1177
2	1024	105	3.56	0.1187
3	1024	100	3.57	0.1190
4	512	110	<b>3.43</b>	<b>0.1143</b>
5	512	105	3.61	0.1203
6	512	100	3.85	0.1283
7	256	110	3.50	0.1167
8	256	105	3.80	0.1267
9	256	100	3.85	0.1283

### 4.3. Effect of combining the IVA and ADReSSo datasets

#### 4.3.1. The ADReSSo regression task results

Since the audio recordings of the two datasets are similar (CT descriptions), we investigated the effect of combining the two datasets on the performance of the regression models. First, we added the IVA dataset to the training set of ADReSSo dataset and repeated the regression tasks. We refer to this model as the ADReSSo (+IVA) model.

Figure 2 shows the RMSE of the nine regression models run using the selected sets of dims and maximum words lengths. As can be seen, two of the models had improvements and the rest either did not change much or got slightly worse. The best ADReSSo (+IVA) model achieved an RMSE of 4.40, which is better than the third winner of ADReSSo 2021 task 2) with dims of 256 and maximum words length of 110 (although considering all results, there are no statistically significant improvements, i.e. p-value > 0.05). Adding data from the other dataset improved the RMSE score by around 0.3. From the experiences with the DNNs, we know that the structure of DNN determines the performance of models and to gain good results the hyper-parameters should be chosen carefully. We also know that the ADReSSo dataset is a challenging dataset and the lack of access to the manual transcriptions did not allow us to train decent ASRs to produce better hypotheses. Also, as mentioned earlier, there is some mismatch between the test set and the training set of the challenge (looking at the results reported by different authors in the challenge, shows that most of the models that performed well on the evaluation set, did not perform comparably on the test set). Here, the data augmentation technique only helped when we selected a good set of hyper-parameters.

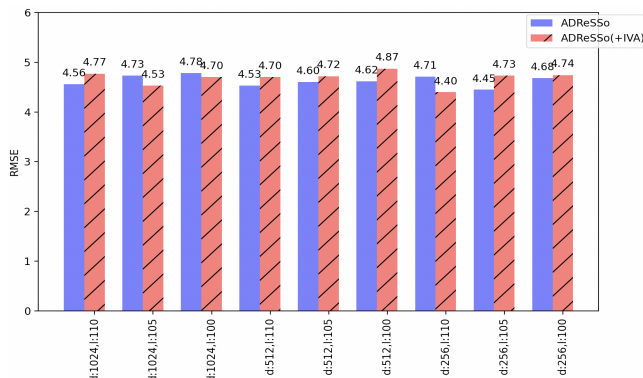


Figure 2: RMSE on different dims (d), maximum length of words (l) comparing ADReSSo vs. ADReSSo (+IVA).

#### 4.3.2. The IVA regression task results

Finally, we added the ADReSSo dataset to the IVA dataset and repeated running the regression models. These are called IVA (+ADReSSo). Figure 3 compares the RMSE of the different models with the selected dims and the maximum word length on IVA and IVA (+ADReSSo). In contrary to ADReSSo (+IVA) models, here, most of the models had significant improvements, and two models achieved the best performance with an RMSE of 3.00 (dims:1024/max-len:105, and dims:256/max-len:110). Combining datasets for the regression modelling task had a better effect on the IVA dataset (these are statistically significant improvements, p-value=0.0178). This might be due to having a better audio quality of recordings and having better ASRs. Also, the ADReSSo dataset has got more diversity in terms of the number of speakers (total number of speakers: 237 vs. 121).

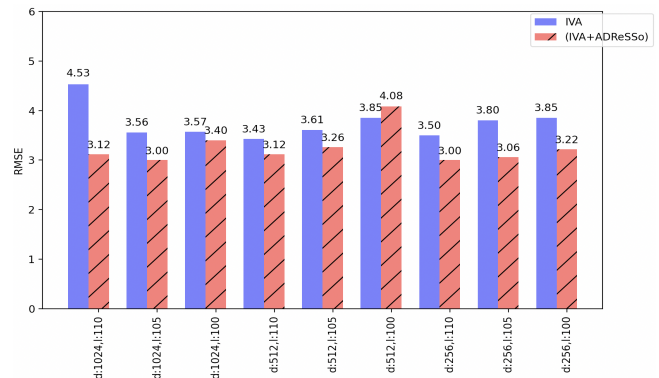


Figure 3: RMSE on different dims (d), maximum length of words (l) comparing IVA vs. IVA (+ADReSSo)

## 5. Conclusions

A key challenge to the robust prediction of cognitive assessment scores is the lack of available data; this is exacerbated by datasets often being accompanied by disparate cognitive scores. In this paper, we explored how the ADReSSo dataset (MMSE scores) might be combined with our in-house IVA dataset (ACE-III scores). We first ported our successful BERT-based classification model introduced in the ADReSSo challenge to the regression task. This model uses different ASR hypotheses and confidence scores which produces multiple inputs resulting in a more robustly trained regression model. The datasets were then combined by converting the ACE-III scores of the IVA dataset to the equivalent MMSE scores. On the ADReSSo regression task, with careful hyper-parameters selection, this improved the performance of the regression model to 4.40 RMSE (in line with a top-three result in the ADReSSo 2021 challenge). On the IVA regression task, even more, significant improvements were achieved with most hyper-parameters. This may be because the ADReSSo dataset, which is acoustically more diverse and with a higher number of speakers, improves robustness for the IVA regression model. Future work will concentrate on further exploring the optimal regression model architecture, as well as exploring more on the confounding effect of ASR and BERT performance.

## 6. Acknowledgements

This work is supported by the Rosetrees Trust and the Stonegate Trust (COMPASS, Grant Agreement No. M934).

## 7. References

- [1] Dementia Statistics, “Deaths due to dementia,” 2018, accessed on October 12, 2019. [Online]. Available: <https://www.dementiastatistics.org/statistics/deaths-due-to-dementia>
- [2] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer’s dementia recognition through spontaneous speech: The address challenge,” *arXiv preprint arXiv:2004.06833*, 2020.
- [3] —, “Detecting cognitive decline using speech only: The address challenge,” *arXiv preprint arXiv:2104.09356*, 2021.
- [4] B. Mirheidari, Y. Pan, D. Blackburn, R. O’Malley, and H. Christensen, “Identifying cognitive impairment using sentence representation vectors,” *Proc. Interspeech 2021*, pp. 2941–2945, 2021.
- [5] B. Mirheidari, D. Blackburn, R. O’Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, “Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2732–2736.
- [6] B. Mirheidari, D. Blackburn, R. O’Malley, A. Venneri, T. Walker, M. Reuber, and H. Christensen, “Improving cognitive impairment classification by generative neural network-based feature augmentation,” *Proc. Interspeech 2020*, pp. 2527–2531, 2020.
- [7] G. Harold and K. Edith, “Boston diagnostic aphasia examination,” *The Psychological Corporation, Philadelphia*, 1972.
- [8] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.
- [9] 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.
- [10] M. Yancheva, K. C. Fraser, and F. Rudzicz, “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*, 2015, pp. 134–139.
- [11] Z. Fu, F. Haider, and S. Luz, “Predicting mini-mental status examination scores through paralinguistic acoustic features of spontaneous speech,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5548–5552.
- [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,” *Proc. Interspeech 2021*, pp. 3810–3814, 2021.
- [13] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velazquez, P. Zelasko, J. Villalba, and N. Dehak, “Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios,” *Proc. Interspeech 2021*, 2021.
- [14] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, “Tackling the addresso challenge 2021: The muer-rmit system for alzheimer’s dementia recognition from spontaneous speech,” *Proc. Interspeech 2021*, pp. 3815–3819, 2021.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [16] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ““mini-mental state”: a practical method for grading the cognitive state of patients for the clinician,” *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.
- [17] S. T. Creavin, S. Wisniewski, A. H. Noel-Storr, C. M. Trevelyan, T. Hampton, D. Rayment, V. M. Thom, K. J. Nash, H. Elhamoui, R. Milligan *et al.*, “Mini-mental state examination (mmse) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations,” *Cochrane Database of Systematic Reviews*, no. 1, 2016.
- [18] A. J. Mitchell, “The mini-mental state examination (mmse): update on its diagnostic accuracy and clinical utility for cognitive disorders,” in *Cognitive screening instruments*. Springer, 2017, pp. 37–48.
- [19] P. Mathuranath, P. Nestor, G. Berrios, W. Rakowicz, and J. Hodges, “A brief cognitive test battery to differentiate alzheimer’s disease and frontotemporal dementia,” *Neurology*, vol. 55, no. 11, pp. 1613–1620, 2000.
- [20] E. Mioshi, K. Dawson, J. Mitchell, R. Arnold, and J. R. Hodges, “The addenbrooke’s cognitive examination revised (ace-r): a brief cognitive test battery for dementia screening,” *International journal of geriatric psychiatry*, vol. 21, no. 11, pp. 1078–1085, 2006.
- [21] S. Hsieh, S. Schubert, C. Hoon, E. Mioshi, and J. R. Hodges, “Validation of the addenbrooke’s cognitive examination iii in frontotemporal dementia and alzheimer’s disease,” *Dementia and geriatric cognitive disorders*, vol. 36, no. 3–4, pp. 242–250, 2013.
- [22] J. A. Matías-Guiu, V. Pytel, A. Cortés-Martínez, M. Valles-Salgado, T. Rognoni, T. Moreno-Ramos, and J. Matías-Guiu, “Conversion between addenbrooke’s cognitive examination iii and mini-mental state examination,” *International psychogeriatrics*, vol. 30, no. 8, pp. 1227–1233, 2018.
- [23] 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 *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.
- [24] J. Glass *et al.*, “Classifying alzheimer’s disease using audio and text-based representations of speech,” *Frontiers in Psychology*, vol. 11, p. 3833, 2020.
- [25] S. Farzana and N. Parde, “Exploring mmse score prediction using verbal and non-verbal cues,” in *INTERSPEECH*, 2020, pp. 2207–2211.
- [26] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, “Using state of the art speaker recognition and natural language processing technologies to detect alzheimer’s disease and assess its severity,” in *INTERSPEECH*, 2020, pp. 2177–2181.
- [27] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, “Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection,” *Proc. Interspeech 2021*, pp. 3790–3794, 2021.