INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Identifying cognitive impairment using sentence representation vectors

Bahman Mirheidari[1], *Yilin Pan*[1], *Daniel Blackburn*[2], *Ronan O'Malley*[2], *and Heidi Christensen*[1]

[1]Department of Computer Science, University of Sheffield, Sheffield, UK
[2]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK
{b.mirheidari,heidi.christensen}@sheffield.ac.uk

Abstract

The widely used word vectors can be extended at the sentence level to perform a wide range of natural language processing (NLP) tasks. Recently the Bidirectional Encoder Representations from Transformers (BERT) language representation achieved state-of-the-art performance for these applications. The model is trained with punctuated and well-formed (writ-ten) text, however, the performance of the model drops significantly when the input text is the – erroneous and unpunctuated– output of automatic speech recognition (ASR). We use a sliding window and averaging approach for pre-processing text for BERT to extract features for classifying three diagnostic categories relating to cognitive impairment: neurodegenerative dis-order (ND), mild cognitive impairment (MCI), and healthy controls (HC). The in-house dataset contains the audio recordings of an intelligent virtual agent (IVA) who asks the participants several conversational questions prompts in addition to giving a picture description prompt. For the three-way classification, we achieve a 73.88% F-score (accuracy: 76.53%) using the pre-trained, uncased base BERT and for the two-way classifier (HCvs. ND) we achieve 89.80% (accuracy: 90%). We further improve these by using a prompt selection technique, reaching the F-scores of 79.98% (accuracy: 81.63%) and 93.56% (accuracy:93.75%) respectively.

Index Terms: Dementia detection, language representation, speech recognition, processing of pathological speech

1. Introduction

Dementia affects the cognitive and communication abilities of people which impairs their speech and language as well as their general ability to undertake daily activities. The number of people living with dementia in the UK is over 850,000 and it is expected to rise to one million by 2025 [1]. Dementia and Alzheimer's is one of the top leading causes of death in the UK [2]. Detecting the early stages of dementia is a challenging task due to overlapping symptoms with normal ageing, limited capability of existing screening tools, and the high risk associated with some procedures such as exposure to radiation. Current cognitive tests routinely assess the speech and communication ability of people but these are inaccurate and difficult to interpret for non-specialists. Therefore, developing automatic assessments methods is of great importance. Recently, we have developed an automatic system analysing people's conversation with an intelligent virtual agent (IVA) to detect cognitive decline [3, 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 description task [5]. In our previous work, we have extracted a number of acoustic, lexical, conceptual and conversational analysis features from the whole conversations between the participants and the IVA, and we showed how the features can be used to reliably distinguish betwee neurodegenerative disorder (ND), mild cognitive impairment (MCI) and healthy controls (HCs) [6, 7, 4]. In this paper, we explore the role of each prompt in the classification accuracy. We use the word vector and language representation (sentence level vectors) to extract the context of the answers given to each question.

Word embedding techniques are widely used in natural language processing (NLP) applications. The early techniques were bag-of-words (BOW) [8] and Frequency Inverse Document Frequency (TF-IDF) [9] which did not consider the order of words nor the context. The next techniques were trained using neural networks such as W2Vec [10] and GloVE [11] which captured the co-occurrences of words and the context of the text. However recently, state-of-the-art performances have been achieved by using transformer-based models like the BERT language model [12]. BERT has been widely used in a variety of natural language processing (NLP) applications like question answering, topic detection, summarisation and semantic search.

Common for those types of applications is that the input text is well-formed and well-punctuated, i.e., matching what the embedding model has been trained on. However, if using BERT in a *speech-driven* pipeline, the input text will no longer be a good match but instead be a continuous string of predicted words. In this paper, we investigate how the automatic speech recognition (ASR) errors and lack of punctuation in the text may affect the performance of BERT model for a system aiming to detect cognitive impairment in spontaneous speech. Using the word vectors and language representation (sentence level vectors) as features, we can train efficient classifiers with high performance to distinguish between different levels of cognitive impairment related to dementia. Using the features we determine a subset of questions that are more useful for the classification (question selection process).

The remainder of the paper contains the following: Section 2 is a brief introduction to related work of using word vector and sentence/text vector techniques. Section 3 contains our experimental setup, especially how we train the ASR, extract the sentence vectors and train the classifiers. Section 4 and 5 cover the results and conclusions.

2. Related work

BERT models have been recently introduced for dementia detection and state-of-the-art results have been reported [13]. There has been a number of studies using BERT in the ADReSS challenge 2020 [14] containing Cookie Theft descriptions, in which the authors shared a training set containing 78 audio files with the corresponding human transcripts of Alzheimer's Disease (AD) participants as well as 78 transcript for non-AD people, and a test set with 48 recordings (24 for each group). The data was taken from the Dementia Bank Corpus [15] to perform two tasks: classification between AD and non-AD, and

http://dx.doi.org/10.21437/Interspeech.2021-915

predicting the mini-mental state examination (MMSE) scores. The successful models in the challenge were mostly trained on the manual transcripts of the recordings. In reality it is not desirable to rely on the manual transcripts as it would be costly and time-consuming to provide them. Instead fully automated systems should be explored, which can take as input the audio recordings. This means they need to be able to deal with the challenges of using ASR systems and combining the automatic transcripts with models such as BERT.

Different groups in the challenge reported different performances. In particular, the models using BERT achieved good accuracies ranging between 75% and 85.4%, e.g. [16] (combining x-vectors and the manual transcripts) reported 75% accuracy; [17] got 81.25% and [18] 83.3%; [19] (adding pauses and disfluency to the transcripts) achieved 85.4% accuracy, similar to the accuracy gained by [20] with a multi-modal (using both audio signals and manual transcripts) BERT.

The ASR errors can affect the word vector techniques significantly. In our previous work, we used GloVe word vectors on 473 recordings of the Dementia Bank dataset and we achieved an accuracy of 75.6% using the manual transcripts without the punctuation. However, the accuracy dropped significantly to 62.3% when we replaced the manual transcripts with the ASR outputs (45.3% WER) [21]. Therefore, we can expect that when using BERT adding the erroneous text from the ASR affects the results considerably. This paper analyses the effect of using ASR transcripts with BERT in a cognitive impairment application and proposes a way of mitigating the effects.

3. Experimental setup

The IVA dataset was collected between 2016 and 2020 at the Department of Neurology, University of Sheffield, UK (Royal Hallamshire Hospital). A total of 168 participants were recorded of which 98 were chosen for this study (61 HC, 19 ND and 18 MCI [1]). The other recordings were only used for training the ASR. The IVA conversation includes nine questions, and the Cookie Theft description task.

3.1. ASR

The LIBRISPEECH dataset was used for training a time delay neural networks (TDNN) acoustic model based on Kaldi's LIBRISPEECH recipe [22]. Then using a 10 fold cross-validation approach, the base acoustic model was adapted to the IVA dataset following the transfer learning technique of [23] (transferring all layers). One epoch of training was carried out on the target dataset (IVA) to adapt the acoustic model. For the language model, the four-gram model was used with Turing smoothing interpolated with the language model of the LIBRISPEECH text. An average 27.9% WER was achieved using the 10-fold cross validation approach.

3.2. Classifiers

Two types of classifiers were chosen for the experiments in this study: the conventional Logistic Regression (LR) classifier and a Transformer based sequence classifiers. The LR classifier is efficient and quick and produces deterministic results, while the Transformer classifiers need longer time to be trained and tuned and each time produce different results depending on the initialisation of the layers' weights. To gain stabilised results, the Transformer classifiers were run five times and then, using a voting approach, the labels were predicted. LR needs a fixed-length feature vector as its input, whereas the Transformers are fed with a fixed-length sequence of vectors.

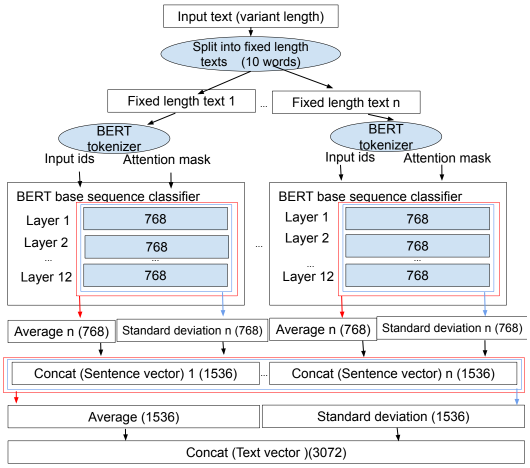
Input text (variant length)
Split into fixed length texts (10 words)
Fixed length text 1
...
Fixed length text n
BERT tokenizer
Input ids
Attention mask
BERT base sequence classifier
Layer 1
Layer 2
...
Layer 12
768
768
...
768
...
BERT tokenizer
Input ids
Attention mask
BERT base sequence classifier
Layer 1
Layer 2
...
Layer 12
768
768
...
768
Average n (768)
Standard deviation n (768)
Average n (768)
Standard deviation n (768)
Concat (Sentence vector) 1 (1536)
...
Concat (Sentence vector) n (1536)
Average (1536)
Standard deviation (1536)
Concat (Text vector)(3072)

Figure 1: *Averaging technique to build the sentence vectors on the the BERT model: each input text is divided into a number of fixed length text (using a sliding window). Then the average of the layers and the standard deviations are combined to build a sentence and whole text vector.*

3.3. Sentence vectors

The Transformer based classifiers can handle input text of varying lengths. If the input text is shorter than a maximum length of words, the spaces are padded with a specific token, while the longer sentences are truncated. BERT models were trained on punctuated corpora. The authors of BERT trained both cased and uncased [2] models of BERT with different sizes of corpora (tiny, small, medium and large). In this study, we use the pre-trained model with a normal size corpus, as well as the large size: uncased BERT base (for convenience we refer to it as BERT), and uncased BERT large (BERT-large).

A sentence can be passed to the pre-trained BERT model as its input and the corresponding sentence vector can be extracted from the weights of the network's layers. However, as mentioned before, the LR classifier works on fixed-length input features. So for BERT with 12 layers and each layer having 768 neurons, per each input word we can extract a vector of $12 \times 768 = 9216$ dimensions. Assume that the maximum word length is 150, then the BERT model can represent the whole input with a vector of $150 \times 9216 = 1,382,400$ dimensions (3.7 million dimensions for BERT-large), which is not feasible to use with an LR classifier. We therefore need to construct a compact version of the sentence/text vectors. To this end, we firstly calculate the average of the vectors as the representation for the sentence. We have tried different approaches and found that averaging the weights on all layers of the model, combined with the standard deviation of the weights (so representation with $2 \times 768 = 1536$ dimensions instead of 1.4 million), was the best; see Figure 1. We also observed that better representation can be achieved by using smaller sentences. So instead of using the whole text we split it into smaller sentences, and us-

[1] These are the numbers that we have collected so far. The numbers of each classes are not balanced, so we did calculate weighted F-score, precision and recall.

[2] dealing with uppercase and lowercase words, respectively.

Table 1: *Two-way classification results on the ADReSS dataset using manual transcript with punctuation (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	80.00	79.17	79.02	79.17
LR (BERT)	**81.75**	**81.25**	**81.18**	**81.25**
LR (BERT-large)	79.37	79.17	79.13	79.17
Transformer (GloVe)	75.17	75.00	74.96	75.00
BERT classifier	82.67	81.25	81.04	81.25
BERT-large classifier	**88.57**	**87.50**	**87.41**	**87.50**

Table 2: *Two-way classification results on the ADReSS dataset using manual transcript without punctuation (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	**84.16**	**81.20**	**80.84**	**81.25**
LR (BERT)	**84.16**	**81.20**	**80.84**	**81.25**
LR (BERT-large)	**84.16**	**81.20**	**80.84**	**81.25**
Transformer (GloVe)	75.00	75.00	75.00	75.00
BERT classifier	**81.75**	**81.25**	**81.18**	**81.25**
BERT-large classifier	78.31	77.08	76.83	77.08

ing the same averaging technique, we calculate the average of those small sentence representations to represent the whole text (i.e. $4 \times 768 = 3072$ dimensions for BERT, $4 \times 1024 = 4048$ demotions for BERT-large, and $4 \times 300 = 1200$ dimensions for GloVe).

If the input text contains punctuation, the whole text can be divided into smaller pieces using the positions of full stops, exclamation or question marks. However, if there is no punctuation (as is the case when working with the output of conventional ASRs) we can split the whole text into a fixed-length sequence of words using the sliding window technique. We found that a window size of 10 and two-word steps for training the LR classifiers resulted in the best best performing classifiers.

4. Results

Before investigating which questions of the IVA dataset are more informative for the classification, it is interesting to first quantify the effect of removing the punctuation and ASR errors. We will do this by investigating this in the ADReSS dataset as well as in our IVA Cookie Theft descriptions.

4.1. Effect of removing punctuation

To demonstrate the effect of the punctuation on the classifiers' performance, we first used the manual transcripts of the ADReSS challenge with and without punctuation by applying the sliding window technique. Table 1 and 2 show the performance of the classifiers in terms of precision, recall, F-score and accuracy measures. Using the sentence word representation based on the punctuation marks, the best performance, achieved by the LR classifier trained on vectors from BERT, was an F-score of 81.18% and an accuracy of 81.25%. However, the best classifier was BERT-large which achieved 87.41% F-score and 87.50% accuracy (comparable to the highest results reported by the ADReSS challenge [14]). Removing the punctuation from the manual text reduced the performance of the BERT-large based classifier (an almost 10% drop to 77.08%). However, the LR classifiers trained on the features, obtained using the sliding window and averaging, were more robust; all reached almost 81% F-score and accuracy. Since we observed more robustness from the LR classifiers than the Transform based ones, we only used the LR classifiers for the following experiments.

4.2. IVA three-way classification

Our IVA contains Cookie Theft picture descriptions that are directly comparable to those in the ADReSS dataset. The IVA has three diagnostic classes though, ND, MCI and HC, and so we first performed the three-way classification tasks. Section 4.3 contains the two-way classifications (HC vs. ND). The tree-way classification task is naturally more difficult than the two-way classification not just because of a lower chance-level, but the classes are also more confusable in terms of how the speech and language is affected with MCI having some resemblance to both HC and ND classes. To show the effect of using the ASR-generated transcripts, we will perform the classification on the manual transcripts then compare them with the classifications using the ASR outputs. Table 3 shows that using the manual transcripts, LR (BERT-large) achieved 72.19% F-score (74.49% accuracy) (LR (BERT) gained a slightly lower F-score). However, as can be seen from the table, using the erroneous ASR transcripts reduced the performance of the three LR classifiers. The best three-way LR classifier was LR (BERT-large) with 68.76% F-score and 71.43% accuracy.

Table 3: *Three-way classification results on IVA Cookie Theft dataset using manual (Man) transcript vs. ASR outputs (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	Man/ASR	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	Man.	72.11	72.45	67.75	72.45
LR (BERT)	Man.	72.78	74.49	72.12	74.49
LR (BERT-large)	Man.	**72.27**	**74.49**	**72.19**	**74.49**
LR (GloVe)	ASR	65.60	66.33	60.99	66.33
LR (BERT)	ASR	66.57	70.41	66.62	70.41
LR (BERT-large)	ASR	**70.26**	**71.43**	**68.76**	**71.43**

4.3. IVA two-way classification

As the next step, the LR classification tasks were repeated on only the two classes (ND vs. HC, which is similar to the ADReSS challenge task). Table 4 shows the LR classifier performances using only the Cookie Theft part of the recordings versus using Cookie Theft and questions. As can be seen given having only Cookie Theft the binary LR (GloVe) classifier outperformed the two other classifiers with 83.60% F-score and 83.75% accuracy. However, adding the nine questions, LR (BERT) achieved 89.80% F-score (90% accuracy).

4.4. Prompt selection

In our previous work, we have always included all of the prompts (questions and various cognitive tasks) given by the IVA. Here we explore whether this is warranted or perhaps different prompts contribute different amounts and so, using only a subset might improve performance. Table 5 shows the results

Table 4: *Two-way classification results on IVA dataset Cookie Theft (CT) using ASR outputs vs. Cookie Theft and all questions (CT + Q's) (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	Prompts	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	CT	**83.47**	**83.75**	**83.60**	**83.75**
LR (BERT)	CT	79.96	81.25	80.20	81.25
LR (BERT-large)	CT	79.33	80.00	79.61	80.00
LR (GloVe)	CT+Q's	87.04	87.50	86.96	87.50
LR (BERT)	CT+Q's	**89.76**	**90.00**	**89.80**	**90.00**
LR (BERT-large)	CT+Q's	89.79	90.00	89.57	90.00

of the three-way classifiers using all questions plus the Cookie Theft text produced by the ASR. As can be seen, adding the text from the questions to the picture descriptions, has improved the performance of the three classifiers. More specifically, LR (BERT) achieved 73.88% F-score and 76.53% accuracy.

Table 5: *Three-way classification results on the IVA dataset (all questions and Cookie Theft) using ASR outputs (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Classifier	Prompts	WPr %	WRc %	WFs %	Ac %
LR (GloVe)	CT+Q's	65.79	69.39	66.58	69.39
LR (BERT)	CT+Q's	**73.76**	**76.53**	**73.88**	**76.53**
LR (BERT-large)	CT+Q's	73.82	75.51	73.13	75.51

For the two types of classification (three-way and two-way) the five most important prompts were selected on the best-performing classifiers from Tables 5 and 4 (we calculated all possible combination of nine questions and Cookie Theft description and then selected the combination with the highest F-score). The results are in Table 6. The prompts selection improved the three-way F-score of the LR (BERT) classifier from 73.88% to 79.98% (accuracy rose from 76.53% to 81.63%). The five best prompts were questions 3 (asking who's most worried about the condition), 4, 5 & 8 (recent memory) in combination with the Cookie Theft prompt. For the two-way scenario, the LR (BERT) classifier achieved 93.56% F-score (93.75% accuracy), and the five best prompts were questions 3 & 6 (distant memory), 7 & 8 (recent memory), and 9 (who manages finances). The two common best prompts amongst the three-way and two-way classifiers were questions 3 and 8.

To show the effect of question selection on the individual classes, confusion matrices of the classifiers were analysed; Figures 2 and 3. For the three-way classification, the question selection slightly decreased the percentage/number of correctly recognised HC participants (from 96.72% (59) to 95.08% (58)), however, the percentage/number of correctly recognised ND increased from 57.89% (11) to 68.42% (13) and for MCI more improvement can be seen from 27.78% (5) to 50% (9). However, the MCI group was the most confused group having more overlap with the other groups. For the two-way classification, the question selection improved the percent/number of correctly recognised ND participants from 68.42%(13) to 78.95%(15), and the number of correctly recognised HC from 96.72%(59) to 98.36% (60).

Table 6: *Prompt (Prmpt) selection for the best three and two way classifiers (Cl.) (Weighted Precision: WPr, Weighted Recall: WRc, WeightedF-score: WFs, Accuracy: Ac).*

Cl.	5 best Prmpt's	WPr %	WRc %	WFs %	Ac %
3-way	**3**, 4, 5, **8**, CT	81.14	81.63	79.98	**81.63**
2-way	**3**, 6, 7, **8**, 9	93.75	93.75	93.56	**93.75**

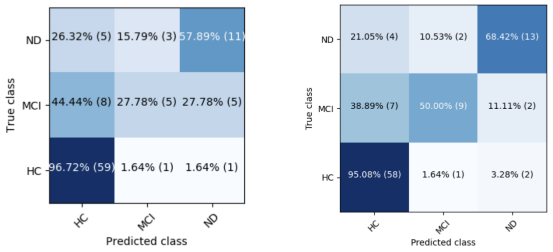

Figure 2: *Confusion matrix of the three-way LR (BERT) classifier using all questions vs. the best 5 questions.*

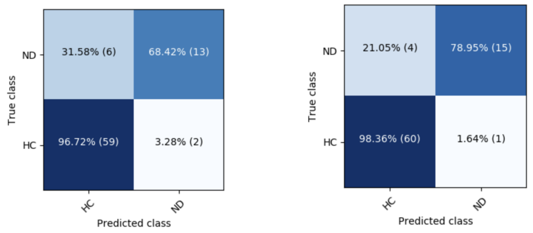

Figure 3: *Confusion matrix of the two-way LR (BERT-Large) classifier using all questions vs. the best 5 questions.*

5. Conclusions

In this paper, we have shown that even though a BERT model can outperform other approaches such as GloVe embedding in many applications, their performance can be affected by the lack of punctuation and the errors introduced when using ASR in fully automated systems. We proposed using a sliding window and averaging technique to produce a sentence representation that can be successfully used as a feature to train a robust LR classifier to identify cognitive decline with high accuracy. Even more improvement was achieved by selecting a subset of the prompts which are more informative for the classification tasks. This demonstrates that the clinically diagnostic conversational questions asked by the IVA should be carefully chosen and that even quite a small set of prompts is capable of giving a very high performance, when used in conjunction with a model like BERT and careful pre-processing of ASR transcripts. For future work, we plan to further investigate technique to mitigate the effect of ASR transcripts on the quality of BERT output.

6. Acknowledgements

This work is supported by the European Union's H2020 Marie Skłodowska-Curie programme (TAPAS; Grant Agreement No. 766287), the Rosetrees Trust and the Stoneygate Trust (COMPASS, Grant Agreement No. M934 and the Fast ASsessment and Treatment in Healthcare funded by EPSRC (Reference. EP/N027000/1).

7. References

[1] D. UK, "What is dementia?" https://www.dementiauk.org, 2021, accessed on March 25, 2021.

[2] Dementia Statistics, "Deaths due to dementia," 2018, accessed on October 12, 2021. [Online]. Available: https://www.dementiastatistics.org/statistics/deaths-due-to-dementia

[3] 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.

[4] 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.

[5] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[6] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "An avatar-based system for identifying individuals likely to develop dementia," *Proc. Interspeech*, pp. 3147–3151, 2017.

[7] B. Mirheidari, Y. Pan, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Data augmentation using generative networks to identify dementia," *arXiv preprint arXiv:2004.05989*, 2020.

[8] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[9] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[11] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[12] 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.

[13] J. Glass *et al.*, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 3833, 2020.

[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] J. T. Becker, F. B. F, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Arch Neurol*, vol. 51, pp. 585–594, 1994.

[16] 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," *Proc. Interspeech 2020*, pp. 2177–2181, 2020.

[17] E. L. Campbell, L. Docio-Fernandez, J. Jiménez-Raboso, and C. Gacia-Mateo, "Alzheimer's dementia detection from audio and language modalities in spontaneous speech."

[18] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[19] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

[20] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *INTERSPEECH (to appear)*, pp. 1–5, 2020.

[21] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations." in *INTERSPEECH*, 2018, pp. 1893–1897.

[22] 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.

[23] V. Manohar, D. Povey, and S. Khudanpur, "Jhu kaldi system for arabic mgb-3 asr challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 346–352.