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

Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-based Alzheimer's Dementia Detection through Spontaneous Speech

Yilin Pan[1*], *Bahman Mirheidari*[1*], *Jennifer M Harris*[2,3], *Jennifer C Thompson*[2,4], *Matthew Jones*[2,4], *Julie S Snowden*[2,4], *Daniel Blackburn*[5], *Heidi Christensen*[1]

[1]Department of Computer Science, University of Sheffield, UK
[2]Division of Neuroscience and Experimental Psychology, University of Manchester, UK
[3]Department of Psychology, University of Exeter, UK
[4]Cerebral Function Unit, Manchester Centre for Clinical Neurosciences, Salford Royal NHS Foundation Trust, Salford UK
[5]Department of Neuroscience, University of Sheffield, UK
{yilin.pan, b.mirheidari, heidi.christensen}@sheffield.ac.uk

Abstract

Exploring acoustic and linguistic information embedded in spontaneous speech recordings has proven to be efficient for automatic Alzheimer's dementia detection. Acoustic features can be extracted directly from the audio recordings, however, linguistic features, in fully automatic systems, need to be extracted from transcripts generated by an automatic speech recognition (ASR) system. We explore two state-of-the-art ASR paradigms, Wav2vec2.0 (for transcription and feature extraction) and time delay neural networks (TDNN) on the ADReSSo dataset containing recordings of people describing the Cookie Theft (CT) picture. As no manual transcripts are provided, we train an ASR system using our in-house CT data. We further investigate the use of confidence scores and multiple ASR hypotheses to guide and augment the input for the BERT-based classification. In total, five models are proposed for exploring how to use the audio recordings only for acoustic and linguistic information extraction. The test results on best acoustic-only and best linguistic-only are 74.65% and 84.51% respectively (representing a 15% and 9% relative increase to published baseline results).
Index Terms: Automatic speech recognition, Alzheimer's dementia, computational paralinguistics

1. Introduction

Research on patients' speech has revealed that linguistic and acoustic abilities are affected even at the early stages of Alzheimer's dementia (AD) [1, 2]. Automatic methods for detecting such impoverishment has focused on extracting acoustic and linguistic information and learning distinguishable patterns for people with and without dementia. For embedding the *linguistic* information, multiple approaches e.g., word2vec [3], hierarchical neural network systems [4], and BERT [5, 6] has been proven to be effective in previous research. For extracting *acoustic* features, both the traditional pipeline systems based on conventional acoustic features [7, 8] and the more recent end-to-end systems for learning task-specific acoustic feature extraction [8, 9] have been explored.

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) challenge is organized for advancing research into automatic AD detection [10]. It contains audio recordings of people describing the Cookie Theft picture. The challenge has not provided manual transcripts of the recordings, in order to reflect more real applications in which it is costly and time-consuming to provide human transcripts. However, automatic approaches to cognitive assessment rely on a combination of linguistic and acoustic information to detect symptoms more comprehensively [11]. To learn linguistic information, the most widespread approach is to first transcribe the audio into text with an automatic speech recognition (ASR) system. Not having manual transcripts of the recordings makes training a good ASR system challenging. To overcome some of these challenges two types of ASR paradigms are explored: i) the traditional ASR systems, based on a pipeline comprising of acoustic, language and lexical models, and ii) end-to-end systems, which directly map a sequence of input acoustic waveforms into a sequence of graphemes or words with an integral structure. For the traditional ASR system, time delay neural network (TDNN) has been used in our previous research and got an outstanding result on dementia-related speech recognition [12–14].

In this paper, BERT is used for extracting linguistic information. However, to mitigate the problems of using ASR-generated transcripts (with erroneous words and lacking punctuation), we explore the use of ASR lattice information. An ASR lattice provides time alignment, recognised words and confidence scores for different hypotheses. Usually, only the most likely hypothesis is selected for the subsequent AD detection [4, 12]. Here, we explore both the use of multiple hypotheses (can be seen as a form of data augmentation) as well as using the hypothesis with the highest word confidence scores (*HighConfHyp*). In addition, we incorporate the confidence scores into the BERT sequence classifier.

Wav2vec2.0 (denoted as *w2v* in the following), as a self-supervised end-to-end ASR system, can achieve similar performance as the traditional ASRs (supervised systems) but with less transcribed audio data. For the w2v structure, information embedded in the transformer layers has been explored in previous research [15, 16]. In our paper, w2v is used both for audio transcription and embedded acoustic feature extraction.

The contribution of the paper can be summarized as follows: (1). Using the acoustic features and automatic transcripts extracted by the w2v results in a superior performance compared to the baseline features reported in [10]. (2). Using mul-

[*] equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-1519

tiple ASR hypotheses and confidence scores as an input to the BERT system improves the performance compared to using just the best hypothesis as in previously proposed approached [12]. (3). Exploring the performance of feature fusion on the acoustic and linguistic information improves results on the provided test set. In the remainder of the paper, Section 2 presents the related work. Section 3 introduces the acoustic and linguistic features used. The experimental setup and the results are discussed in Section 4 and Section 5. Conclusions are given in Section 6.

2. Related work

In the ADReSS Challenge, the predecessor to the ADReSSo challenge, the organisers provided both the acoustic recordings and the corresponding manual transcripts. The best performances were achieved by [6] and [5] using acoustic-only and linguistic-only systems, respectively. Among the 34 participating teams, 7 out of 13 Interspeech-2020 papers used BERT for linguistic-based modelling, thus demonstrating the efficiency of BERT when manual transcripts are available. However, for fully automatic systems, the transcription task is handled by an ASR which introduces word-level errors. It has been found that fine-tuning using such noisy text data, can negatively impact the performance of BERT [17]. Our paper explores how to increase the performance of BERT-based classification when working with ASR transcripts.

Disfluencies and unclear pronunciation can decrease the performance of ASR systems, whilst at the same time be beneficial if the related information is used to inform the classification [5, 12]. In [5], the pause and disfluency annotation was used for punctuation generation providing important linguistic information in addition to the manual transcripts. In [12], ASR-transcribed words with low confidence scores were removed from the generated transcripts to successfully guide the linguistic information extraction. The time alignment information from the ASR output is used for designing the rhythm features for assisting the extracted acoustic features.

Table 1: *Analysis of HighConfHyp on HC and AD.*

parameters (mean&var)	**AD**	**HC**
word duration (s)	$.129 \pm (.013)$	$.120 \pm (.013)$
pause duration (s)	$.267 \pm (.589)$	$.126 \pm (.154)$
confidence score	$.867 \pm (.036)$	$.886 \pm (.033)$
#words/transcript	$65 \pm (1910)$	$81 \pm (3207)$

The lack of manual transcripts in the challenge means that it is difficult to train an ASR system tailored to the specifics of ADReSSo. It also makes it more challenging to evaluate the outputs of any transcripts produced for the ADReSSo data. To get around this, confidence scores are used as a proxy measure for accuracy, essentially replacing the information provided by monitoring WERs for different ASR systems during development. To assess the meaningfulness of using this approach, we analysed the transcripts produced by our system [12] when run on the ADReSSo dataset. Table 1 shows that the mean and variance of pause duration in the AD group was longer than the HC (healthy control) group. Likewise, fewer words were recognised in the AD group compared to the HC group. This might be because people living with AD tend to speak less and with more disfluencies. Moreover, a significant number of the words pronounced by people living with AD are typically not clear enough to be recognised correctly by the ASR as demonstrated in our previous work [12].

3. Acoustic and linguistic features

In this section, two ASR paradigms are introduced for transcribing the audio recordings for linguistic information extraction using BERT. The w2v, is also used for acoustic information extraction. The feature fusion is implemented on the ASR transcripts and the extracted acoustic features.

3.1. Automatic speech recognition

To transcribe the audio recordings into texts, we trained a conventional Kaldi-based ASR which produces decoding lattices allowing us to construct different hypotheses (extracted text from different paths in the lattice). For each hypothesis it is possible to calculate the confidence scores of the words, reflecting how confident the ASR is in recognizing each word.

Table 2: *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.**
DR INTERVIEWS (295)	64.3h	39.2k	736	5.9s
IVA (168)	26.7h	8.3k	219	11.5s
HALLAM (54)	26.14h	10.5k	139	4.8s
SHEFMAN CT (238)	3.9h	0.2k	238	11.5s
LIBRISPEECH (281241)	961.1h	281.2k	5466	12.3s
AMI (682)	95.5h	133.9k	171	2.6s

Since the manual transcripts of the training set of the challenge were not available and using any part of Dementia Bank not permitted, training a high performance ASR tuned to recognise spontaneous speech, ideally of people describing the Cookie Theft picture, was challenging. LIBRISPEECH is a well-known dataset containing almost 1000 hours of audio recordings of people reading books. It was used to build a base TDNN ASR following Kaldi's LIBRISPEECH recipe [18]. Since the dataset is read speech, a transfer learning technique was applied to adapt to spontaneous speech data using a number datasets: (AMI [19], DR INTERVIEWS, IVA, and HALLAM. Table 2 shows information such as length, number of utterances and speakers of the datasets. DR INTERVIEWS , and HALLAM are two datasets collected locally at the Sheffield Royal Hallamshire hospital. They contain audio interviews between neurologists and people with seizure/non-epileptic attacks, and dementia or other memory issues, respectively. In addition, the IVA dataset (also in-house) contains conversation between patients and an Intelligent Virtual Agent. Moreover, 238 Cookie Theft description from Sheffield and Manchester universities (SHEFMAN CT) were used for a second round of adaptation. Of these, a small subset of 20 out of 238 samples was held out for ASR testing and the rest were added to the other datasets for transfer learning. Following [20] (using both the structure and weights of the base ASR and then running one epoch of the DNN model to adapt to the new datasets) the acoustic model of our ASR was constructed. To train the language model, the four-grams with Turing smoothing was applied on the training set. Decoding on a held out set of the SHEFMAN CT data (10%) resulted in a WER of 8.23%.

3.2. Acoustic feature extraction

W2v as an end-to-end ASR paradigm was used for both the audio transcription and the acoustic feature extraction. W2v encodes raw wave $\mathcal{X}$ with multiple Convolutional Neural Networks (CNNs) into latent representations $Z \in z_1, ..., z_T$ for T time-steps. Before passing the inputs to the Transformer to get the textualized representation, first Z was fed to a quantization module [21] for masking [22–24]. The model was first trained on the unlabeled data and then fine-tuned on a small labelled dataset with a Connectionist Temporal Classification (CTC) loss [25, 26]. Research has shown that the representations underlying pre-trained w2v can capture the speaker features and language features [15] embedded in the acoustic recordings. The contextualized representations of the input raw wave were built into the 24 layers transformer architecture. For extracting the acoustic features from the w2v structure, the transformer layers' outputs were extracted as the acoustic representations of the input waveform segments. For each hidden layer output, the extracted hidden feature matrix $[N * feat\_dim]$ was averaged across the length N of the feature. At the same time, the transcribed words for the waveform were used for the following linguistic-based information modelling.

The pre-trained model[1] (named as pre-trained w2v for convenience) is adapted on the IVA dataset (see Table 2) for a better performance on the ADReSSo data. The IVA recordings were split into 60/20/20 parts training, evaluation and testing. After data adaptation, the WER decreased from 31.9% to 18.6% on this IVA test set. The ASR transcripts are used as the input of BERT for linguistic information modelling.

3.3. Combining linguistic and acoustic features

The acoustic features described in Section 3.2 and the linguistic features extracted from the last transformer layer in BERT by concatenating with a fully connected layer are combined as follows. The acoustic features $\mathcal{V} \in [N * feat\_dim]$ were averaged over feature numbers into a vector $v_1 \in feat\_dim$. After dimension reduction with a fully connected layer, the feature v_1 was concatenated with the BERT last-second layer output feature v_2 for fine-tuning.

3.4. Using ASR hypotheses and confidence scores

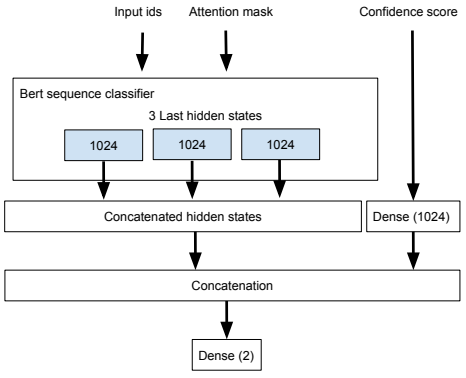

Figure 1: *The last three states of the BERT sequence classifier are concatenated with the confidence score input, before the last classification output layer.*

A range of ASR hypotheses can be derived from the decoder lattices using different parameters: the language weights were between 6 and 10 with a word insertion penalty of 0, 0.5 and 1 producing 30 different hypotheses for each recording. The average confidence scores of the words were calculated for each of the hypotheses. Two models (with different maximum word length of the input sequence: 105 and 100) were built using the uncased BERT large sequence classifier and hypotheses with the confidence scores. Figure 1 shows the structure of the models which was simply a concatenation of the last three hidden states of the BERT classifier with the input confidence score layer. The combination of these two layers was passed finally to the output layer to classify between AD and HC transcripts. Using a variety of outputs from the ASR (which we know are erroneous) alongside the corresponding confidence scores, could help the network to be trained more robustly on the words produced by the ASR.

4. Experimental setup

Table 3 shows the list of the five models submitted to the challenge. Model 1 and Model 2 correspond to the acoustic features and ASR transcripts as output by the w2v model described in 3.2, while Model 3 corresponds to the feature fusion proposed in Section 3.3. For these three models, 10-fold cross-validation (CV) was applied[2]. Then the results (Section 5) were averaged across the 10 folds, and the test labels were estimated by the majority voting on the predict labels from the 10 folds trained models. However, for Model 4 and Model 5 corresponded to Section 3.4, since it was a time-consuming process to run multiple fold-based evaluations, instead a single evaluation set was constructed by holding out 20% of the training set.

4.1. Model setup

BERT-for-Sequence-Classification[3] [27] was used for modelling the linguistic information. Two configurations of BERT models were used with a transformer layer inside the models of 12 layers (BERT$_{base}$) and 24 layers (BERT$_{large}$). Two dense layers were added to the BERT$_{base}$ for feature fusion with 256 dimensions. To fine-tune the BERT$_{base}$ and BERT$_{large}$ using the ASR-derived transcripts of the ADReSSo training set, the parameters were set as in Table 3. For fine-tuning the w2v, 168 IVA recordings, shown in Table 2, were used.

4.2. Model selection

We used our evaluation set to select the five models submitted to the challenge. As part of this process, we also evaluated a number of proposed models based on our and other's previous work. Using the approach in [5], the performance of the HighConfHyp on BERT$_{base}$ using 10 fold averaged accuracy was 72.77% on our evaluation set. Also, we replicated the experiments in [5] on the BERT3p model (the best performance with BERT in the paper) based on the HighConfHyp. The time alignment information was used for generating the punctuation insertion instead, and the 10-fold CV averaged accuracy on the evaluation set was 75.06%. Inspired by the experiments in [12], the transcribed words in the HighConfHyp are selected with the confidence scores by replacing the word with the confidence score lower than 0.87 (selected based on the average confidence score of the two classes in Table 1) by <*unk*>. The 10-fold CV

[1] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

[2] nine folds of training set for training and one fold for evaluation

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

averaged accuracy on the evaluation set is 76.65%. The evaluation result on the pre-trained w2v transcripts is 78.01% under the same parameters as in Model 2.

Table 3: *AD detection models, TB=Tree Bagger, CS=confidence scores, model parameter: e=epochs, mwl=max word length, ne=number of estimators, bs=batch size]*

Alias	Information Input	Classifier	Parameters
Model 1	w2v acoustic feat.	TB	ne=10
Model 2	fine-tuned w2v ASR	$\text{BERT}_{base.}$	mwl=512; e=8;bs=4
Model 3	w2v outputs fusion	BERT_{base}	mwl=512; e=8;bs=4
Model 4	ASR hypotheses+CS	BERT_{large}	mwl=105; e=1; bs=64
Model 5	ASR hypotheses+CS	BERT_{large}	mwl=100; e=1; bs=64

5. Results

5.1. Acoustic feature comparison

For classifying the extracted acoustic features, four linear classifiers were selected, namely decision trees (DT), nearest neighbour (KNN), TB and support vector machines (SVM). The acoustic features are normalized before classifying. The acoustic features' evaluation results corresponding to the 24 transformer layers extracted from the pre-trained w2v model and fine-tuned model are shown in Figure 2 respectively. As shown, the highest accuracy (80.08%) was achieved by the TB classifier with the 16th hidden layer features extracted from the pre-trained model; this was selected as model 1.

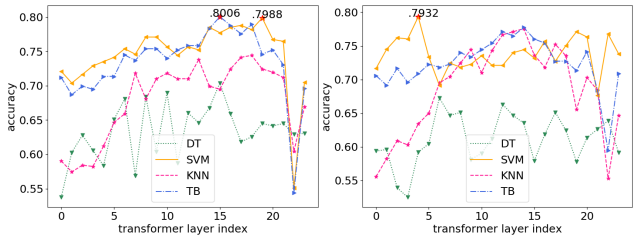

Figure 2: *Comparing classifiers using acoustic features extracted from the wav2vec2-large-960h-lv60 pre-trained model (left) and fine-tuned model (right).*

5.2. Evaluation results

The baseline results on the leave-one-subject-out (LOSO) evaluation (CV) set reported by the authors of the challenge [10] were 78.92% and 72.89% on the acoustic and linguistic-based models respectively. The results corresponding to our five models are listed in Table 4. All models outperform the CV baseline[4]. In particular, the feature fusion result on the evaluation set (83.47%) from model 3 is better than the acoustic-only result (80.06% for model 1) and linguistic-only (81.59% for model 2), though not as good as the multiple ASR hypotheses and confidence scores corresponded models (model 4 and model 5), which achieved accuracies of 97.39% and 96.69% respectively.

[4]A direct comparison on evaluation data is difficult, as we have not been able to evaluate using LOSO because of time constraints and have instead evaluated using 10-fold CV (models1, 2 and 3) or on 20% held out data (models 4 and 5).

Table 4: *Final classification results on the evaluation set. Pr: Precision, Rc: Recall, Fs: F-score,Ac: Accuracy*

Model	Pr %		Rc %		Fs %		Ac %
	HC	AD	HC	AD	HC	AD	
Model 1	80.69	81.04	78.39	81.17	78.72	80.50	80.06
Model 2	83.63	83.02	79.82	83.08	79.32	82.25	81.59
Model 3	86.25	86.33	83.57	83.08	82.80	82.65	83.47
Model 4	**98.69**	**96.26**	**95.78**	**98.85**	**97.22**	**97.54**	**97.39**
Model 5	98.46	95.19	94.51	98.66	96.45	96.90	96.69

Table 5: *Final classification results on the test set. Pr: Precision, Rc: Recall, Fs: F-score,Ac: Accuracy*

Model	Pr %		Rc %		Fs %		Ac %
	HC	AD	HC	AD	HC	AD	
Model 1	72.50	77.42	80.56	68.57	76.32	72.73	74.65
Model 2	76.19	**86.21**	**88.89**	71.43	82.05	78.13	80.28
Model 3	76.92	81.25	83.33	74.29	80.00	77.61	78.87
Model 4	85.29	81.08	80.56	85.71	82.86	83.33	83.10
Model 5	**87.88**	81.58	80.56	**88.57**	**84.06**	**84.93**	**84.51**

5.3. Test results

The baseline results on the test set were 64.79% and 77.46% for the acoustic and linguistic systems, respectively [10]. The final classification results of the five models we proposed are shown in Table 5. The acoustic-only and best linguistic-only results achieved 74.65% and 84.51% accuracy respectively (models 1 a& 5), which all outperforms the baseline models. Interestingly, the feature fusion based model (model 3) performed better than model 1 and model 2 on the evaluation set, but not as well as the linguistic-only models on the test set. This might indicate a mismatch between the evaluation set and the test set.

6. Conclusion

In our paper, two ASR paradigms were adopted for linguistic and acoustic feature extraction. For modelling the linguistic information, multiple ASR hypotheses and confidence scores were passed to the pre-trained BERT for model tuning on the ASR transcripts from the ADReSSo training set. The acoustic features were extracted from the transformer outputs of the pre-trained w2v model. The BERT model based combination of the acoustic and linguistic information improved the performance of the classifier on the evaluation set, but not on the test set. In the future, the combination between the acoustic information and the multiple ASR hypotheses is expected to be explored to improve the test set performance.

7. 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 Trus (COMPASS, Grant Agreement No. M934 and the Fast ASsessment and Treatment in Healthcare funded by Engineering and Physical Science Research Council (EPSRC) (Reference. EP/N027000/1).

8. References

[1] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[2] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in Alzheimer's disease, is that an early sign? importance of changes in language abilities in Alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

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

[4] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic hierarchical attention neural network for detecting ad." in *INTERSPEECH 2019*. ISCA-International Speech Communication Association, 2019, pp. 4105–4109.

[5] 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," *INTERSPEECH 2020*, pp. 2162–2166, 2020.

[6] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for Alzheimer's dementia through spontaneous speech," *INTERSPEECH 2020*, pp. 1–5, 2020.

[7] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[8] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," in *INTERSPEECH 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

[9] Y. Pan, B. Mirheidari, Z. Tu, R. O'Malley, T. Walker, A. Venneri, M. Reuber, D. Blackburn, and H. Christensen, "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification," *INTERSPEECH 2020*, pp. 4806–4810, 2020.

[10] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," *medRxiv*, 2021.

[11] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[12] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of Alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," in *INTERSPEECH 2020*. ISCA-International Speech Communication Association, 2020, pp. 4961–4965.

[13] 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," in *INTERSPEECH 2020*. ISCA-International Speech Communication Association, 2020, pp. 2527–2531.

[14] 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.

[15] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[17] A. Srivastava, P. Makhija, and A. Gupta, "Noisy text data: Achilles' heel of BERT," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, 2020, pp. 16–21.

[18] 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.

[19] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88. Citeseer, 2005, p. 100.

[20] 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.

[21] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[22] 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.

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

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[25] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *arXiv preprint arXiv:1911.03912*, 2019.

[26] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6