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

Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios

Raghavendra Pappagari[1], *Jaejin Cho*[1], *Sonal Joshi*[1], *Laureano Moro-Velazquez*[1], *Piotr Żelasko*[1,2], *Jesús Villalba*[1,2], *Najim Dehak*[1,2]

[1]Center for Language and Speech Processing, Johns Hopkins University, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{rpappag1, jcho52, sjoshi12, laureano, pzelasko, jvillal7, ndehak3}@jhu.edu

Abstract

In this study, we analyze the use of speech and speaker recognition technologies and natural language processing to detect Alzheimer disease (AD) and estimate mini-mental status evaluation (MMSE) scores. We used speech recordings from Interspeech 2021 ADReSS$_o$ challenge dataset. Our work focuses on adapting state-of-the-art speaker recognition and language models individually and later collectively to examine their complementary behavior for the tasks. We used speech embedding techniques such as x-vectors and prosody features to characterize the speech signals. We also employed automatic speech recognition (ASR) with interpolated language models to obtain transcriptions used to fine-tune the BERT models that classify and assess the speakers. Our results indicate that the fusion of scores obtained from the multiple acoustic and linguistic models provides the best detection results, suggesting that they contain complementary information. A separate analysis of the models indicates that linguistic models outperform acoustic models in detection and prediction tasks. However, acoustic models can provide better results than linguistic models under certain circumstances due to the errors in ASR transcriptions, which indicates that the performance of linguistic models relies on the performance of ASRs. Our best models provide 84.51% accuracy in automatic detection of AD and 3.85 RMSE in MMSE prediction.

Index Terms: Alzheimer Disease, Automatic Speech Recognition, Mini-Mental Status Evaluation

1. Introduction

The most common signs of Alzheimer disease (AD)[1], are memory decline, disorientation, confusion, and behavior changes. This leads to loss of independence, having a clear impact on patients, their families, and the society [2]. The prevalence of AD and related dementias in the USA for populations older than 65 years old is 11.5%, with an increasing incidence due to the improvement in life expectancy in the coming decades, which would double the associated burden by 2060 [3].

While two of the most typical signs of AD are memory and cognitive decline, language impairment is also common, as it is linked to cognitive and memory-related problems and neurodegenerative processes. In this respect, speech technologies can deliver new precision medicine tools that will provide an objective quantitative analysis and reliable proof, analysis, comparison, and circulation for a faster diagnosis.

The literature suggests some common signs in the speech of AD patients related to articulatory aspects such as apraxia of speech [4] or others linked to communication and word retrieval deficits such as progressive, logopenic, or anomic aphasia [5, 6], or anomia [4]. In this respect, pause and silence-related features allow characterizing the loss of verbal fluency, which is associated with AD [7]. These problems of verbal fluency are caused, in part, by the difficulties that patients have in recalling words, finding the appropriate vocabulary, or finishing sentences. The use of verbal fillers such as /um:/, or /eh:/ or the description of a word instead of the use of that word, are also common. In more advanced stages, stuttering, repetition of ideas and questions, and difficulties forming simple sentences become frequent [7].

In the Interspeech 2020 ADReSS challenge [8], numerous teams proposed different approaches to detect AD and automatically predict mini-mental status evaluation (MMSE)[2] in a dataset containing speech and manual transcriptions from 78 AD patients and 78 sex and age-matched controls. Whereas some of the participant teams focused on using either speech or linguistic approaches, the results from several teams indicate that approaches containing a combination of different linguistic aspects and, in some cases, acoustic aspects lead to better results [9, 10, 11, 12, 13], providing detection accuracy over 75% in the evaluation subset. Some authors employed term frequency–inverse document frequency (TF-IDF) features such as grammatical dependency and universal dependency features [9, 14], with different classifiers such as XGBoost or logistic regression. Other authors used a transformer-based pre-trained language model (LM) based on bidirectional encoder representations from transformers (BERT) [10, 15, 16, 14, 17, 13, 18], and other neural network approaches such as bi-directional Hierarchical Attention Networks [11] or Transformer XL [13]. All of the linguistic approaches used the manual transcriptions provided by the challenge organizers, and none of them analyzed the use of any automatic speech recognition (ASR) system to obtain transcriptions using audio in the detection or regression tasks.

Approaches using acoustic modeling involved the use of x-vectors [10, 17], i-vectors [17], bag of audio words [11], spectral and cepstral features with different classifier backends [9, 16, 11], acoustic features obtained with OpenSMILE [8, 19, 20, 13], and VGGish [13, 18] with heterogeneous results that, in general, did not outperform linguistic approaches.

In this study, we propose the use of several acoustic and linguistic models to detect and assess AD for the Interspeech 2021 ADReSS$_o$ challenge [21]. The main difference from the 2020

[1]The possessive form has been deliberately removed in this article, following the World Health Organization and the US National Institutes of Health recommendations [1].

[2]MMSE ranges between 0 and 30 and is used to assess the dementia status of patients, being values higher than 24 considered as normal cognition.

http://dx.doi.org/10.21437/Interspeech.2021-1850

challenge is that the manual transcriptions of the audio recordings were not included in the dataset. Thus, we employed ASR systems with adapted LMs, and commercial ASR platforms to obtain transcriptions from the audio and used linguistic models using the obtained text. At the same time, we employed acoustic models using several acoustic representations derived from state-of-the-art speaker and speech recognition technologies as well as prosodic features to detect and assess AD automatically.

The code of the experiments is being shared by the authors of this paper[3].

2. The ADReSS$_o$ challenge

In this paper, we addressed two of the three tasks proposed by the challenge organizers:

- *AD detection task* - automatic differentiation between participants with and without AD using a short speech session.
- *MMSE prediction task* - automatic prediction of the participant's MMSE using the same dataset of the AD detection task.

2.1. The Dementia Bank-ADReSS$_o$ 2021 dataset

The ADReSS$_o$ challenge dataset, described in [21], contains the *diagnosis dataset* with speech from speakers with and without AD. The recordings include a picture description, employed for AD detection and MMSE regression tasks. In most cases, these recordings consist of a participant's interaction with one investigator under several recording conditions with different types of background noise. The dataset is divided into *training* and *evaluation* subsets. The *training* subset contains 87 recordings from speakers with AD and 79 from control subjects and the *evaluation* subset, a total of 71 recordings.

3. Methods

In this study, we analyzed multiple acoustic and linguistic modeling approaches to carry out the automatic AD detection and MMSE prediction tasks proposed in Section 2. Then, we performed a score-level fusion of these approaches to obtain new predictions, as indicated in the following sections. There were two types of experiments:

- Cross-validation: performed by training and testing with the *training subset*, using a 10-fold scheme where class distributions were consistent over the folds.
- Evaluation: obtained by testing the models trained on the *training subset* with the *evaluation subset*. For each separate approach, we propagated the *evaluation subset* through an ensemble classifier that averages the scores from the 10 cross-validation models.

3.1. Acoustic modeling

We used several types of acoustic modeling to characterize the speech from the dataset and represent the speakers' articulatory, prosodic and phonatory traits. On the one hand, we used an end-to-end classifier by fine-tuning an x-vector model [22]. On the other hand, we extracted different types of embeddings and acoustic features from available libraries as input to logistic regression and XGBoost classifiers.

[3]https://github.com/sonal-ssj/ADReSSo_2021_JHU

3.1.1. x-vectors

An x-vector model is a deep neural network that generates one single vector (embedding) per recording, characterizing the full signal. Although the technique is considered the current state-of-the-art for speaker recognition, several studies suggest that these embeddings also contain information related to emotion, speaking rate, gender [23, 24] and can be used to characterize the influence of neurological diseases, such as Parkinson disease on speech [25]. The x-vector architecture considered in this study is the same as the employed in [26], and contained three main parts: an encoder network to extract frame-level representation from Mel-frequency cepstral coefficient (MFCC), a global temporal pooling layer to produce the embedding (x-vector), and a feed-forward classification network to produce speaker class posteriors. For the encoder, we used a ResNet-34 [27] structure consisting of a sequence of 2D convolutional layers with residual connections between them. The pooling network comprises a multi-head attention layer and operates on the ResNet output. Different heads are designed to capture different speech aspects of the input signal. We concatenated the attention heads output and pass it through a fully connected layer whose output is passed through an utterance-level classifier to obtain model decision.

We pre-trained this model for speaker recognition using VoxCeleb1, VoxCeleb2, NIST SRE4-10, and Switchboard datasets similarly as in [10]. Then, we replaced the last part of the model, the fully connected layer, to detect AD using softmax in the output or provide MMSE values using linear activation in the output depending on the task, and retrained the whole model. Additionally, we trained a second model in the same way but with noise and music augmentation (x-vectors augm), as indicated in [26] to obtain more robust representations. Both types of models (x-vectors and x-vectors augm) were trained considering two different frame-lengths: 25 and 250 ms. We note that as x-vector models were pre-trained for speaker classification, our models could perform well on AD detection using attributes related to speaker classification instead of using AD characteristics. By evaluating on unseen subjects, however, we made sure that our model's performance is reflective of its ability to capture AD characteristics.

3.1.2. Embeddings and prosody features

Encoder-decoder ASR embeddings. We computed embeddings using the encoder of an encoder-decoder ASR system included in the SpeechBrain library [28]. We used an acoustic model trained on LibriSpeech [29], that consists of an encoder with convolutional recurrent deep neural networks (CRDNN) architecture followed by a bidirectional LSTM and a fully connected layer to obtain the acoustic representation that we call as "encoder-decoder ASR embeddings" (SB Enc/Dec).

Prosody features. Previous studies have found that temporal features of AD patients differ from those of controls as the patients tend to have more silent pauses than controls [7]. We used DigiPsychProsody [4] to compute prosody features. These included total speech time, total pause time, percentage pause time, speech pause time, mean pause duration, and pause variability. These features are computed using 3 different intensities of the WebRTC[5] Voice Activity Detector. We obtained these features: (1) for the entire conversation recordings (2) per speaker – investigator and patient.– In this last case, we used the

[4]https://github.com/NeuroLexDiagnostics/DigiPsych_Prosody
[5]https://github.com/wiseman/py-webrtcvad

segmentation files given by the organizers to separate speech for each speaker. When we used segmentation, we concatenated the prosody features obtained from investigator and patient.

VGGish features. VGGish is a feature embedding front-end for audio classification models that has provided good results in AD detection [13, 18]. We used a pre-trained model [6] trained using the AudioSet dataset [30].

eGeMAPS Features. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features are a selected standardized set of statistical features that characterize affective physiological changes in voice production. We extracted these features for the entire recording as we expect them to capture overall speaker characteristics.

Embeddings and prosody features classification. The feature vectors obtained with SpeechBrain, the prosody features extractor, VGGish and eGeMAPS were employed to carry out the different challenge tasks in combination with logistic regression and XGBoost classifiers. All possible combinations of representations and classifiers were evaluated in cross-validation.

3.2. Linguistic modeling

3.2.1. Automatic speech recognition

Since the challenge data does not include human-annotated transcripts, we used ASR models to transcribe the recordings. As the recordings contain conversational speech and has noticeable noise and reverberation – with the microphones often being located far from speakers – we used a pre-trained ASpIRE recipe model[7] in Kaldi [31]. ASpIRE was a far-field speech recognition challenge held by IARPA, and the model in the recipe is trained on the English portion of the Fisher corpus [32], which is conversational speech, with data augmentation through room impulse response convolution and background noise [33]. This model is denoted as ASR 1 in our experiments.

Then, to improve the transcription quality, we interpolated the ASpIRE LM with an LM trained on automatic transcripts of the target domain (*training subset* of ADReSS$_o$ 2021) obtained with ASR 1. This interpolated model will provide more likelihood to frequent words of the target domain, leading to a lower word error rate. This system is denoted as ASR 2.

Lastly, to obtain other automatic transcripts from ASR systems trained on more varied acoustic data and possibly cover multiple linguistic domains, we employed two commercially available ASR systems: Amazon Web Services (AWS) and the Otter.ai ASR models denoted as ASR 3 and ASR 4, respectively. Since challenge data does not have manual transcriptions, we could not compare how different ASRs perform in regard to word error rate (WER) on the data.

3.2.2. BERT language model

We modeled the linguistic-phonological manifestations of AD using a pre-trained LM, BERT [34], on the automatic transcriptions of the speech recordings. BERT has provided state-of-the-art performances in multiple applications such as question answering, natural language inference, sentence, and word prediction, sentiment prediction, among many others [35]. After fine-tuning BERT, the transformer-based model can be used to model language context, flow, and complexity in the tasks of interest of this study [10]. In similar lines, previous works have shown promising results in other tasks such as depression detection [36] and sentiment analysis [37].

The BERT architecture consists of self-attention layers and feed-forward layers, similar to transformer encoder layers. The inputs of the model were tokens from the automatic transcript using WordPiece tokenizer [38]. The input token sequence was processed through the multiple encoder layers until the penultimate layer to obtain embeddings for each token. Then, the sequence of token embeddings was pooled to pass through a last linear layer to obtain the final prediction. In our case, we used a pre-trained BERT model and adapted it to our tasks (AD detection and MMSE prediction) in the following manner:

- We replaced the last layer of the model with a task-specific layer: a linear layer having two outputs with a softmax activation function for AD detection or a linear layer having one output with a linear activation function for MMSE prediction.
- We fine-tuned the entire pre-trained model using our data to minimize the cross-entropy loss for AD detection or mean square error for MMSE prediction.

For each iteration of the cross-validation experiments, 8 folds from the *training subset* were employed for BERT fine-tuning, 1 fold for early stopping, and the remaining fold for testing. We fine-tuned the models for up to 5 epochs.

3.3. Model fusion

We explored model fusion by using the output scores of the models as the inputs of a logistic regression classifier to obtain a final prediction (detection or MMSE prediction). We first obtained fused models by combining the acoustic models. Then, we combined the best acoustic model and the best linguistic model, As for linguistic modeling, we used just BERT with different ASR transcriptions. Finally, we combined the best fused acoustic model with the best linguistic model, i. e., we first fused several acoustic models, and the resulting scores were fused with those from BERT.

4. Results and discussion

Acoustic models. Table 1 contains the results of the different acoustic models for the detection and MMSE prediction using logistic regression as a classifier.[8] Detection results are reported using accuracy (%) and MMSE prediction results, using root mean square error (RMSE).

Results in the first block of Table 1 indicate that all of the acoustic features provide some differentiation between classes. SB Enc/Dec embeddings with logistic regression and x-vectors model with 250 ms frame-length provide the best cross-validation results for AD detection, whereas SB Enc/Dec embeddings and eGeMAPS provide the best RMSE values for MMSE prediction. Prosody features related to pause times and pause vs. speech ratios characterize the loss of verbal fluency, which is associated with AD [7] and, whereas these provide only 61% accuracy, the results suggest that these help to automatically differentiate between speakers with and without AD.

Linguistic models. The second block in Table 1 includes the accuracy and RMSE results of four BERT models fine-tuned with automatic transcriptions obtained with four ASR systems. The best detection result is obtained from the one fine-tuned with the automatic transcriptions from ASR 3, and best MMSE prediction with ASR 4, both of which are based on commercially available ASR systems. These results suggest that transcription errors lead to worse prediction and detection results

[6] https://github.com/tensorflow/models/tree/master/research/audioset, https://github.com/harritaylor/torchvggish

[7] https://kaldi-asr.org/models/m1

[8] We also used other classifiers like XGBoost, however, since logistic regression outperformed other classifiers, due to space constraints, we have included results for only logistic regression.

Table 1: *Best AD detection accuracy (%) and MMSE prediction RMSE using acoustic (top), linguistic (middle), and fusion (bottom) modeling during cross-validation.*

Model	Detection accuracy(%)	MMSE RMSE
Acoustic		
x-vectors	69.3	7.18
x-vectors (250 ms)	71.1	6.97
x-vectors augm	58.4	6.92
x-vectors augm (250 ms)	63.9	7.01
VGGish	60.8	6.92
SB Enc/Dec	**71.7**	**6.44**
Prosody (20 ms)	60.8	6.94
Prosody (30 ms)	59.0	6.89
Prosody per speaker (20 ms)	61.5	6.96
Prosody per speaker (30 ms)	61.5	7.05
eGeMAPS	63.3	6.76
Linguistic		
BERT (ASR 1)	63.1	6.74
BERT (ASR 2)	69.5	5.90
BERT (ASR 3)	**76.1**	5.44
BERT (ASR 4)	73.3	**5.39**
Fusion		
All acoustic models	69.9	6.74
x-vector (250 ms), SB Enc/Dec	72.3	6.55
x-vector, x-vector (250 ms), SB Enc/Dec	73.5	6.60
x-vector, x-vector (250 ms), SB Enc/Dec, Prosody (20 ms)	72.9	6.76
BERT (ASR 3), {x-vector, x-vector (250 ms), SB Enc/Dec }	**81.3**	**5.23**

using linguistic approaches. Whereas in our previous work [10], linguistic models trained with manual transcriptions always outperformed acoustic models, in this study, some acoustic models from Table 1 outperform those linguistic models that possibly have high word error rates in their transcriptions. On the other hand, there is a remarkable improvement in the results of the linguistic model with ASR 2 transcription compared to with ASR 1 transcription. This indicates that interpolating the LM in ASR with the automatic transcription from the in-domain recordings can improve the final linguistic modeling. Therefore, new linguistic-based diagnostic tools will benefit from LM interpolation when the speech tasks (cognitive tests) are known.

Score-level fusion. Results of score fusion are included in the third block of Table 1. All of these values are reported using logistic regression as the fusion back-end, since it led to the best results. The fusion of the scores of all acoustic models led to worse results than the fusion of only two or three acoustic models. For score fusion with only acoustic models, the score fusion of SB Enc/Dec with several modalities of x-vector models showed the best result with 73.5% accuracy in cross-validation. However, the fusion of the acoustic models does not reduce the RMSE in the prediction task. The fusion of the linguistic model trained with the automatic transcriptions from ASR 3 and three acoustic models provides the best cross-validation results of the study, 81.3% and lowest RMSE in prediction, 5.23. This coincides with the findings of past work [10] that suggests that acoustic and linguistic approaches can have complementary information for detection and assessment of AD.

Evaluation results. Following the challenge rules, we submitted the scores from five different models on the detection task and other five on the MMSE task for challenge evaluation. Table 2 includes the results of that evaluation in terms of precision, recall, F1-score, and accuracy for detection, and RMSE in the prediction task. The trend of the results is similar to those in the cross-validation. Notably, the linguistic approach employing the ASR 3 transcription and its fusion with acoustic approaches provide the best results in terms of accuracy. However, in the evaluation result, the linguistic approach employing the ASR 4 transcriptions provides the best MMSE prediction. In general, the evaluation results are better than those obtained using cross-validation. One possible reason can be the use of ensemble models, built with the 10 cross-validation models for each of the submitted approach for evaluation, prevents overfitting.

Table 2: *ADReSS$_o$ challenge evaluation results for the detection and prediction tasks. Best results are marked in bold. Ac. fusion refers to the fusion of scores from acoustic models x-vector, x-vector (250 ms) and SB Enc/Dec. Global fusion refers to the fusion of Ac. fusion scores with BERT (ASR 3) scores. Following the challenge rules, 5 models were submitted for evaluation of detection, and other 5 models for prediction tasks. * indicates that the system was not submitted for evaluation*

		Detection			*Prediction*
Model	Class	Rec/Prec	F1	Accu (%)	RMSE
Baseline [21]	CC	0.78/0.80	0.78	78.87	5.28
	AD	0.80/0.78	0.78		
SB Enc/Dec	CC	0.72/0.72	0.72	71.80	5.74
	AD	0.71/0.71	0.71		
Ac. fusion	CC	0.75/0.75	0.75	74.70	*
	AD	0.74/0.74	0.74		
BERT (ASR 2)	CC	0.92/0.77	0.84	81.70	4.67
	AD	0.71/0.89	0.79		
BERT (ASR 3)	CC	0.94/0.79	0.86	**84.51**	4.26
	AD	0.74/0.92	0.83		
BERT (ASR 4)	CC	*	*	*	**3.85**
	AD	*	*		
Global fusion	CC	0.94/0.79	0.86	**84.51**	4.62
	AD	0.74/0.92	0.83		

5. Conclusions and future work

In this study, we have analyzed the use of acoustic and linguistic approaches for the automatic detection of AD and MMSE prediction in a low resource scenario proposed by the ADReSS$_o$ challenge organizers. The acoustic approaches consisted of speaker and speech recognition embeddings and prosodic features, whereas the linguistic models were built with BERTs trained on different ASR transcripts. Our findings suggest that acoustic and linguistic approaches contain complementary information for automatic detection and assessment of AD. The x-vector model and encoder-decoder automatic speech recognition embeddings provided the best results among acoustic models, and the BERT fine-tuned with automatic transcriptions from a commercial ASR system yielded the best results for the linguistic approach. Also, the use of the interpolated LM to adapt the ASR to the target domain produced an absolute improvement of 6.4% accuracy and 0.84 in the detection and MMSE prediction tasks, respectively.

In future work, we will evaluate the use of several iterations of LM interpolation to adapt and refine the ASR to the target domain. We will also explore multi-modal approaches in which the classifier uses aligned linguistic and acoustic information in order to extract more precise cues and exploit the bi-modal complementarity.

6. References

[1] K. Ayesu, B. Nguyen, S. Harris, and S. Carlan, "The case for consistent use of medical eponyms by eliminating possessive forms," *Journal of the Medical Library Association: JMLA*, vol. 106, no. 1, p. 127, 2018.

[2] K. B. Rajan, J. Weuve, L. L. Barnes, R. S. Wilson, and D. A. Evans, "Prevalence and incidence of clinically diagnosed alzheimer's disease dementia from 1994 to 2012 in a population study," *Alzheimer's & Dementia*, vol. 15, no. 1, pp. 1–7, 2019.

[3] K. A. Matthews, W. Xu, A. H. Gaglioti, J. B. Holt, J. B. Croft, D. Mack, and L. C. McGuire, "Racial and ethnic estimates of alzheimer's disease and related dementias in the united states (2015–2060) in adults aged 65 years," *Alzheimer's & Dementia*, vol. 15, no. 1, pp. 17–24, 2019.

[4] E. Rochon, C. Leonard, and M. Goral, "Speech and language production in alzheimer's disease," *Aphasiology*, vol. 32, no. 1, pp. 1–3, 2018.

[5] J. D. Rohrer, M. N. Rossor, and J. D. Warren, "Alzheimer's pathology in primary progressive aphasia," *Neurobiology of aging*, vol. 33, no. 4, pp. 744–752, 2012.

[6] S. M. Harnish, "Anomia and anomic aphasia: Implications for lexical processing." *The Oxford Handbook of Aphasia and Language Disorders*, 2018.

[7] S. H. Ferris and M. Farlow, "Language impairment in alzheimer's disease and benefits of acetylcholinesterase inhibitors," *Clinical interventions in aging*, vol. 8, p. 1007, 2013.

[8] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *Proc. Interspeech 2020*, pp. 2172–2176, 2020.

[9] M. Martinc and S. Pollak, "Tackling the adress challenge: a multimodal approach to the automated recognition of alzheimer's dementia," *Proc. Interspeech 2020*, pp. 2157–2161, 2020.

[10] R. Pappagari, J. Cho, L. Moro-Velázquez, 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.

[11] N. Cummins *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.

[12] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech," in *Proc. Interspeech*, 2020, pp. 2187–2191.

[13] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting multimodal features from pre-trained networks for alzheimer's dementia recognition," *Proc. Interspeech 2020*, pp. 2217–2221, 2020.

[14] T. Searle, Z. Ibrahim, and R. Dobson, "Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech," *Proc. Interspeech 2020*, pp. 2192–2196, 2020.

[15] 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.

[16] 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," *Proc. Interspeech 2020*, pp. 2167–2171, 2020.

[17] A. Pompili, T. Rolland, and A. Abad, "The inesc-id multi-modal system for the adress 2020 challenge," *Proc. Interspeech 2020*, pp. 2202–2206, 2020.

[18] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *Proc. Interspeech 2020*, pp. 2222–2226, 2020.

[19] E. Edwards *et al.*, "Multiscale system for alzheimer's dementia recognition through spontaneous speech," *Proc. Interspeech 2020*, pp. 2197–2201, 2020.

[20] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity," *Proc. Interspeech 2020*, pp. 2212–2216, 2020.

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

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[23] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," *arXiv preprint arXiv:2002.05039*, 2020.

[24] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[25] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[26] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," *ICASSP 2020*, p. (accepted), 2020.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] M. Ravanelli *et al.*, "Speechbrain," https://github.com/speechbrain/speechbrain, 2021.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[30] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[31] D. Povey, A. Ghoshal, and G. Boulianne, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[32] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[33] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms." in *ASRU*, 2015, pp. 539–546.

[34] 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.

[35] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4593–4601.

[36] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[37] S. Pei, L. Wang, T. Shen, and Z. Ning, "Da-bert: Enhancing part-of-speech tagging of aspect sentiment analysis using bert," in *International Symposium on Advanced Parallel Processing Technologies*. Springer, 2019, pp. 86–95.

[38] M. Johnson *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.