Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data

Ning Wang[1], *Yupeng Cao*[1], *Shuai Hao*[1], *Zongru Shao*[*2 3] , *K.P. Subbalakshmi*[1]

[1]Stevens Institute of Technology, NJ, Hoboken
[2]Center for Advances Systems Understanding, Görlitz, Germany
[3]Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany

nwang7@stevens.edu, ycao33@stevens.edu, shao8@stevens.edu, drssth@gmail.com, ksubbala@stevens.edu

Abstract

In this work, we propose a modular multi-modal architecture to automatically detect Alzheimer's disease using the dataset provided in the ADReSSo challenge. Both acoustic and text-based features are used in this architecture. Since the dataset provides only audio samples of controls and patients, we use Google cloud-based speech-to-text API to automatically transcribe the audio files to extract text-based features. Several kinds of audio features are extracted using standard packages. The proposed approach consists of 4 networks: C-attention-acoustic network (for acoustic features only), C-Attention-FT network (for linguistic features only), C-Attention-Embedding network (for language embeddings and acoustic embeddings), and a unified network (uses all of those features). The architecture combines attention networks and a convolutional neural network (C-Attention network) in order to process these features. Experimental results show that the C-Attention-Unified network with Linguistic features and X-Vector embeddings achieves the best accuracy of 80.28% and F1 score of 0.825 on the test dataset.
Index Terms: Alzheimer's disease, Multi-Modal Approach, CNN-Attention network, Acoustic feature, Linguistic feature

1. Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease that is the most common form of dementia and continual cognitive impairments [1]. The number of cases is increasing rapidly every year so that AD has become a non-negligible social public health problem. Therefore, early diagnosis of AD is an essential task and has attracted much attention in recent years.

The ADReSSo challenge at INTERSPEECH 2021 defines a shared task through which different approaches to target the automatic detection of AD [2] can be proposed. The ADReSSo challenge provides only audio data of patients extracted from the Pitt Corpus [3].

Our approach uses both the audio features directly extracted from the audio files and linguistic and other language-based features extracted from the transcribed version of the same audio file. Literature suggests that speech impairment is a common and significant sign of AD even at the early stage of dementia [4, 5]. Therefore, some speech characteristics, such as speech vagueness and abnormal pauses, can function as an important bio-marker. These features in patients' speech can provide useful information about the cognitive status and other aspects related to the level of brain health [6]. Further, studies have shown that several lexical or syntactic features and increases in conversational fillers or non-specific nouns are also indicators of AD [7, 8]. Consequently, Natural Language Processing (NLP) methods can be applied to extract linguistic features from text data [9] and used in the detection of AD [10, 11]. Existing AD classification methods can be divided into three categories depending on the types of features used: leveraging raw audio data or acoustic features using linguistic features derived from text or a combination of acoustic features and linguistic features to detect AD. We have taken the third approach here. The main contributions of this work are as follows: 1) a CNN-attention Network (C-Attention Network) for automated detection of Alzheimer's disease. 2) a method to integrate features extracted from both text and audio.

2. Related Work

Automated detection of Alzheimer's disease has a long history of research. In early automated AD detection work, researchers attempt to quantify the impairments by using computational methods [12]. They first construct or extract different features from the different data sources and then apply traditional machine learning methods to detect Alzheimer's disease. These features can be divided into two categories: linguistic features and acoustic features. Linguistic features including part-of-speech (POS) tag frequencies, measures of lexical diversity were extracted and a linear discriminant analysis or other classifiers were used to identify AD patients [12, 13]. Acoustic features such as mel-frequency cepstral coefficients (MFCC) and low-level descriptors (LDD) were used in [14]. Then, the combination of both acoustic and linguistic features based machine learning approaches were proposed to automatically detect AD [15, 16]. These studies have shown that the methods of combining different types of features have better detection accuracy compared to using features separately.

In recent related research work, the INTERSPEECH 2020 ADReSS challenge provides a baseline paper, which summarized many useful acoustic features [17], including emboase [18], ComParE [19], eGeMAOS [20] and MRGG [21] and followed it with machine learning methods such as Linear Discriminant Analysis (LDA), Decision Tree (DT), Support Vector Machine (SVM) and Random Forests (RF) to detect AD. In ADReSS 2020 challenge, the work [22] utilized two acoustic features, IS10-Paralinguistics feature set from ComParE and Bag-of-Acoustic-Words (BoAW), to achieve a good classification accuracy [22]. Cummins et.al proposed an end-to-end convolutional neural network to directly classify AD [23]. Pan et.al considered the problem of audio data quality and they applied ASR techniques to identify high-quality speech segments

*This work was done when this author was at Stevens Institute of Technology

for more robust feature extraction to improve detection performance [24]. Some researchers obtain latent features from language embeddings and used the attention mechanism to achieve better performance on text data [6, 9, 25]. A multi-modal approach that fused acoustic and linguistic features was proposed in [26]. In that work, the author used dual-LSTM architecture, one for audio feature and another for text feature. A gating mechanism was used to fuse the two for the final classifier.

3. Proposed Approach

In this section, we introduce the acoustic and linguistic feature sets we use and propose a modular multi-modal architecture to classify AD from non-AD controls.

3.1. Acoustic and Linguistic Feature Sets

3.1.1. Acoustic Features

We used open source audio processing toolkits, OpenSMILE [18] and Kaldi [27], to obtain four different acoustic features from the raw audio file, which are Emobase [18], IS10 [19], VGGish [28] and X-Vector [29]. Specifically, Emobase and IS10 are frame-level acoustic features. VGGish and X-Vector are acoustic embeddings. Frame-level features are directly extracted from audio files and these features capture the frequency characteristics and other statistical information. Different from frame-level features, embedding features are not directly derived from the audio data. The embedding features are from the embedding model, where the embedding model will generate a vector to represent the characteristics in audio data. The embedding model is a deep neural network and pre-trained on large audio datasets. We used these pre-trained embedding models to extract features. Here are the specific descriptions for different feature extraction processes:

Emobase: The Emobase feature set has abundant audio features which include mel-frequency cepstral coefficients (MFCC) information, fundamental frequency (F0), F0 envelope, line spectral pairs (LSP), and intensity features.

IS10: The IS10 feature set includes many frame-level features: 16 types of LLDs, PCM loudness, eight log Mel frequency band (0-7), eight line spectral pairs (LSP) frequency (0-7), F0 envelope, voicing probability, jitter local, jitter DPP, and shimmer local and more MFCC features.

VGGish: This is an acoustic embedding model which is pre-trained on YouTube's Audio dataset [28]. The architecture of VGGish is a CNN-based structure and similar to VGG. The VGGish embedding model extracts and transforms the audio features into semantic and meaningful high-level feature vectors with 128 dimensions.

X-Vector: X-vector is a deep neural network-based audio embedder, widely used in the field of speech recognition [29, 30]. We employ x-vector to represent audio features from raw audio files. The neural network that produces the x-vector consists of three components: the frame-level layers to extract representation from MFCC, a statistics pooling layer which receives output from the last frame-level layer and a segment-level layer that follows the statistics pooling layer to generate the x-vector. Specifically, we obtain the x-vector features according to the following steps: 1) First, all raw audio files are normalized and re-sampled to 16,000Hz and 16-bits by using SOX audio processing software; 2) Second, we compute the x-vector for each audio segment by using Kaldi that uses the SER16 pre-trained x-vector model. The SER16 pre-trained model is trained on Switchboard, Mixer 6, and NIST SERs datasets [29, 30]; 3) Third, we convert x-vector to a binary file to make it easier for our proposed model to read.

3.1.2. Features from Transcribed Text

We used Speech-to-Text API [1] provided by Google cloud to automatically transcribe speech recordings. Then based on the transcripts, we extracted linguistic features and sentence embeddings.

Linguistic Features: We used two tools to generate linguistic features: 1) Like [2], we converted transcripts into CHAT format, then ran EVAL and FREQ commands in CLAN [31] to generate a composite profile of 34 measures and Moving Average Type Token Ratio [32]; 2) we generated 22 Part-of-Speech tags using NLTK [15]. After removing all-zero and duplicate features, 50 linguistic features in total were extracted.

Sentence embeddings: We used Universal Sentence Embedding (USE) [33] to represent each sentence in the context.

3.2. Proposed Architectures: C-Attention Networks

We propose a modular multi-modal architecture consisting of three standalone networks. The architectures are shown in Fig 1. The left-hand side leg processes acoustic features, such as Emobase and IS10 features, and is called C-Attention-Acoustic Network. The middle leg processes linguistic features, and is called C-Attention-FT Network. The right-hand side leg processes embedding features, such as USE, VGGish and X-Vector, and is called C-Attention-Embedding Network.

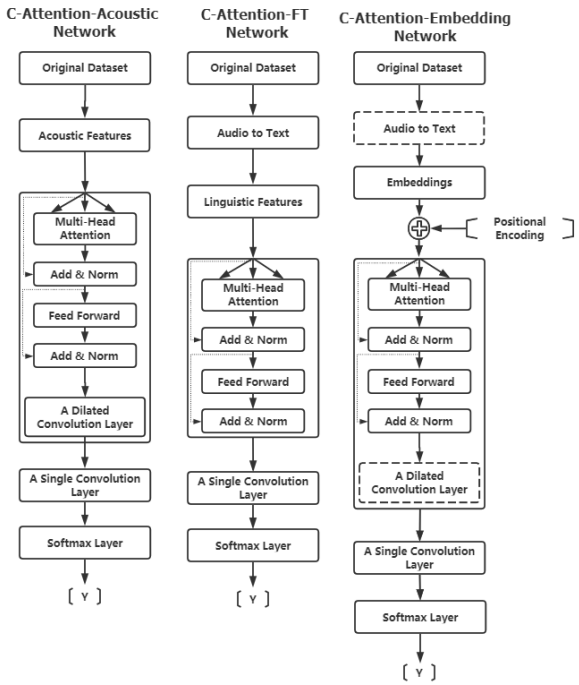

Figure 1: *The proposed architecture of C-Attention-Acoustic Network, C-Attention-FT Network and C-Attention-Embedding Network. The C-Attention-Acoustic Network uses acoustic features, the C-Attention-FT network uses the linguistic features, and the C-Attention-Embedding network uses embeddings of the patient/control's recordings.*

[1] https://cloud.google.com/speech-to-text

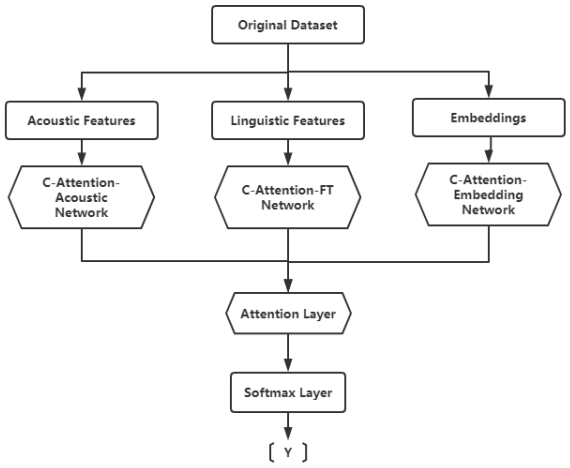

Figure 2: *The Architecture of Unified C-Attention Network for Acoustic Features, Linguistic Features, and Embeddings*

3.2.1. C-Attention Acoustic Model

This architecture (C-Attention Acoustic Network) is depicted on the left-hand side of Figure 1. The C-Attention Acoustic Model comprises of a multi-head-attention (MHA) module together with a dilated convolution layer [34, 35]; followed by a 1-D CNN layer and a softmax layer. We used the same MHA module and the encoder structure of the transformer that was proposed in [36]. Let $R = \{r_1, r_2, ..., r_n\}$ be the set of speech recordings, then r_i is the i^{th} record in the dataset. We extract acoustic feature sets presented in Sec 3.1.1 and generate the acoustic feature vectors, Let $F = \{F_1, F_2, ..., F_n\}$ be the set of acoustic feature vectors, and F_i be the i^{th} vector in the acoustic matrix. The MHA transforms the feature matrix $F = \{F_1, F_2, ..., F_n\}$ to another matrix of n-dimensional vectors $A = \{A_1, A_2, ..., A_n\}$. After each MHA module, we use a dilated convolution layer to further distill the MHA matrix $A = \{A_1, A_2, ..., A_n\}$ to half its original size. This is done to reduce the dimensions of the acoustic features which are too large for the attention mechanism to capture interactions well. This procedure forwards from j^{th} layer into $(j+1)^{\text{th}}$ layer as

$$\mathrm{X_{j+1}} = \mathrm{MaxPool}\left(\mathrm{ELU}\left(\mathrm{Conv1d}\left(\mathrm{X_j}\right)\right)\right) \tag{1}$$

Where the Conv1d(·) performs a 1-D convolutional filters and ELU(·) [37] is the activation function. The MHA and dilated CNN module is followed by a 1-layer CNN and a softmax layer to get the final classification.

3.2.2. C-Attention FT Model

This architecture (C-Attention FT Network) is depicted in the middle of Figure 1. It is proposed to capture the interaction among linguistic features. This architecture is similar to the proposed C-Attention (Sec 3.2.1) except for the removal of dilated CNN layer.

3.2.3. C-Attention Embedding Model

This architecture (C-Attention Embedding Network) is depicted on the right-hand side of Figure 1. We propose this architecture as a means of capturing latent feature information implicit in embeddings. This architecture is similar to the proposed C-Attention (Sec 3.2.1) except for the addition of a positional encoding module. The positional encoding module is used to maintain the relative positions of the embedding features and is the same as that used in the transformer [36] architecture. More specifically, the Audio to Text layer is only applied to text embeddings and the dilated convolution layer is only used on X-Vector embeddings.

3.2.4. C-Attention Unified Model

This architecture (C-Attention Unified Network) is depicted in Figure 2. In this architecture, we use all three types of features: acoustic features, linguistic features, and embedding features. We used another attention layer to fuse the outputs from C-Attention Acoustic Network, the C-Attention-FT network, and the C-Attention-Embedding network followed by a softmax layer. In order to fuse these other models together, we omit the final softmax layers in each of the four modules.

4. Experiments

In this work, we employed four models on acoustic features, linguistic features, and embeddings to detect AD, and evaluated our proposed models on the ADReSSo challenge dataset.

4.1. Dataset

In this work, we employed four models on acoustic features, linguistic features, and embeddings to detect AD, ane dataset is a balanced sub-dataset of the DementiaBank [38] with respect to age and gender. It consists of spontaneous speech recordings of spoken picture descriptions elicited from participants through the Cookie Theft picture description in the Boston Diagnostic Aphasia Exam [39]. The training set consists of 166 speech recordings, including 87 speech recordings from AD patients and 79 speech recordings from healthy controls. The testing set consists of 71 speech recordings without annotations.

4.2. Experiment Setup

We implemented our proposed models using Pytorch and trained them using the 10-fold cross-validation (CV) approach. Three types of features were extracted: acoustic features, linguistic features, and embeddings (including text embeddings and acoustic embeddings). For all models implemented in this paper, each model has 6 multi-head attention layers. Apart from that, in the C-Attention-Acoustic network and C-Attention-Embedding Network, each multi-head attention module is followed by a dilated convolution layer (kernel width=3) and a max-pooling layer with stride 2 which downsizes the feature set into its half slice. We found that due to the variation of feature size, the best configuration of modules is slightly different among various feature sets. For Emobase and IS10 features, 6 multi-head attention modules and 6 dilated CNN modules gave the best performance. However, 6 multi-head attention modules plus two dilated CNN modules is the best setting for X-Vector embeddings. Dilated CNN modules were not used on VGGish embeddings.

4.3. Feature Generation

We have described how to generate each type of feature in Sec 3.1.1. Here we add a few additional explanations on acoustic features and acoustic embeddings used in our experiments.
Acoustic Features: We generated Emobase and IS10 on each

speech recording, no segmentation was applied.

Acoustic Embeddings: 1) **VGGish Embeddings**: We applied 16k-downsampling on single-channel audio signals, and computed the log mel spectrogram. Then each log mel spectrogram was segmented with a non-overlapping 960ms window. Finally, we generated 128-length VGGish embedding on each segmented sample; 2) **X-Vector embeddings**: Similarly, we segmented each speech recording with a non-overlapping 960ms window, then generated 512-length X-Vector embedding on each segmented sample.

4.4. Experiment Results

The performance results are shown in Table 1. We note that on the training dataset, the C-Attention-Unified model with Linguistic and X-Vector features achieved the best performance in respect to the accuracy, precision, and F1 score, the best Recall was achieved by the C-Attention-Unified model with Linguistic, IS10, and X-Vector. Due to time limitation, in C-Attention-Unified network, we were not able to use all feature sets, such as USE. Given C-Attention-Embedding (USE) did not perform better than other approaches but it took longer to train, we truncated this feature set in our unified model.

Table 1: *AD classification accuracy on 10-fold cross-validation (CV)*

Approach	Accuracy	Precision	Recall	F1
C-Attention-Acoustic(Emobase)	0.614	0.632	0.632	0.632
C-Attention-Acoustic(IS10)	0.62	0.615	0.736	0.67
C-Attention-FT(Linguistic)	0.735	0.753	0.736	0.735
C-Attention-Embedding(USE)	0.657	0.679	0.655	0.667
C-Attention-Embedding(VGGish)	0.735	0.759	0.724	0.741
C-Attention-Embedding(X-Vector)	0.753	0.774	0.747	0.76
C-Attention-Unified(Linguistic + USE)	0.711	0.714	0.747	0.73
C-Attention-Unified(Linguistic + VGGish)	0.747	0.771	0.736	0.753
C-Attention-Unified(Linguistic + X-Vector)	**0.772**	**0.787**	0.74	**0.763**
C-Attention-Unified(Linguistic + IS10 + X-Vector)	0.725	0.724	**0.778**	0.75

Our experiment results would indicate that: 1) using both audio embeddings and linguistic features seems to be the best way to approach the problem of detecting AD, rather than choosing only one; 2) On the text side, handcrafted linguistic features perform better than USE representations on AD detection; 3) However, on the audio side, audio embeddings, such as X-Vector and VGGish show better performance on AD detection than frame-level acoustic features, such as Emobase and IS10.

Further analysis of the values in this table would indicate that using only the latent NLP-based features does not perform as well as using only audio embeddings (X-Vector). However, it is worthy to mention that the transcripts used in this work were automatically converted from speech recordings. The automatic conversion might have introduced errors and noises. Within audio embeddings, X-Vector performs better than VGGish.

As part of the ADReSSo challenge, we were provided the test dataset and asked to submit the labels from five attempts of our algorithm on this dataset. Since we had multiple models, we used the following method to decide on which model's result to report. We randomly split the training dataset into 80% for training, 10% for validation, and 10% for testing. We tried multiple random seeds, then used the models which performed best, on an average, on the training dataset. The best performing model was the C-Attention-Unified model with Linguistic features and X-vectors. Hence we used this model to submit the five attempts on the test dataset as required by the challenge rules. The organizers then calculated the accuracy, precision, recall and F scores based on the ground truth labels (which were not revealed to the participants) of the test dataset. Table 2 shows the results returned to us by the organizers, for our model.

Table 2: *Attempts on test dataset*

Attempts	Accuracy	Precision		Recall		F1	
Attempt 1	**0.8028**	0.7500	0.8889	0.9167	0.6857	**0.8250**	0.7742
Attempt 2	0.7746	0.7500	0.8065	0.8333	0.7143	0.7895	0.7576
Attempt 3	0.7887	0.7692	0.8125	0.8333	0.7429	0.8000	0.7761
Attempt 4	0.7746	0.7273	0.8519	0.8889	0.6571	0.8000	0.7419
Attempt 5	0.7606	0.7111	0.846	0.8889	0.6286	0.7901	0.7213

5. Future Work

In this challenge, due to time limitation, we were not able to apply segmentation on acoustic features, nor apply 100ms window size segmentation on either VGGish or X-Vector embeddings. We believe that our models could learn better on the acoustic features if time series segmentation is applied. Further, we will continue to address the other subtasks set out in the challenge, viz.: evaluate the models' performance by calculating the MMSE score and generalize the proposed models to predict the cognitive decline.

6. Conclusions

In this paper, we proposed a modular multimodal approach to detect Alzheimer's disease and this approach includes four architectures using CNN and multi-head attention on the training set of the ADReSSo Challenge. Three types of feature sets were used in this work: acoustic features, linguistic features, and embeddings. One architecture uses only the acoustic features, one architecture uses only the linguistic features, one uses only the embeddings and the unified architecture uses all of those features. Extensive experimental evaluations on the training dataset show that our proposed model can detect AD with an accuracy of 77.2%, F1 of 0.763 using the C-Attention-Unified model with Linguistic and X-Vector features. Using the same model and feature sets, the best accuracy of our models was 80.28% and F1 of 0.825 on the test dataset.

7. References

[1] 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.

[2] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo Challenge," in *Submitted to INTERSPEECH 2021*, 2021. [Online]. Available: https://edin.ac/31eWsjp

[3] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[4] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013.

[5] 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.

[6] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, "Using state of the art speaker recognition and natural language pro-

cessing technologies to detect alzheimer's disease and assess its severity," *Proc. Interspeech 2020*, pp. 2177–2181, 2020.

[7] P. Garrard, L. M. Maloney, J. R. Hodges, and K. Patterson, "The effects of very early alzheimer's disease on the characteristics of writing by a renowned author," *Brain*, vol. 128, no. 2, pp. 250–260, 2005.

[8] V. Berisha, S. Wang, A. LaCross, and J. Liss, "Tracking discourse complexity preceding alzheimer's disease diagnosis: a case study comparing the press conferences of presidents ronald reagan and george herbert walker bush," *Journal of Alzheimer's Disease*, vol. 45, no. 3, pp. 959–963, 2015.

[9] 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.

[10] N. Wang, M. Chen, and K. P. Subbalakshmi, "Explainable CNN-attention networks (c-attention network) for automated detection of alzheimer's disease." ACM SIGKDD, 2020. [Online]. Available: https://arxiv.org/pdf/2006.14135.pdf

[11] N. Wang, F. Luo, R. Chandramouli, K. P. Subbalakshmi, and V. Peddagangireddy, "Personalized early stage alzheimer's disease detection: A case study of president reagan's speeches." ACL, 2020.

[12] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[13] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech," in *IEEE International Conference Mechatronics and Automation, 2005*, vol. 3. IEEE, 2005, pp. 1569–1574.

[14] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] 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.

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

[17] 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.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[19] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[20] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[21] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[22] 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.

[23] 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.

[24] 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," *Proc. Interspeech 2020*, pp. 4961–4965, 2020.

[25] 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.

[26] 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.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[28] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[29] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[30] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[31] B. MacWhinney, "Tools for analyzing talk part 2: The clan program," *Pittsburgh, PA: Carnegie Mellon University. Retrieved from http://talkbank. org/manuals/CLAN. pdf*, 2017.

[32] M. A. Covington and J. D. McFall, "Cutting the gordian knot: The moving-average type–token ratio (mattr)," *Journal of quantitative linguistics*, vol. 17, no. 2, pp. 94–100, 2010.

[33] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[34] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[35] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *arXiv preprint arXiv:2012.07436*, 2020.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[37] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[38] F. Boller and J. Becker, "Dementiabank database guide," *University of Pittsburgh*, 2005.

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