DEMENTIA DETECTION BY FUSING SPEECH AND EYE-TRACKING REPRESENTATION

Zhengyan Sheng[1], *Zhiqiang Guo*[1], *Xin Li*[1,2], *Yunxia Li*[3], *Zhenhua Ling*[1]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China
[2]iFlytek Research, Hefei, P.R.China
[3]Shanghai Tongji Hospital, Tongji University School of Medicine, Shanghai, China
$\{sa21006159, gzq\}$ $@mail.ustc.edu.cn$, $\{zhling, leexin\}@ustc.edu.cn$, $doctorliyunxia@163.com$

ABSTRACT

This paper proposes a method of detecting dementia from the simultaneous speech and eye-tracking recordings of subjects in a picture description task. First, automatic speech recognition (ASR) and regional picture recognition (RPR) models are built to extract content-related bottleneck (BN) features for both speech and eye-tracking inputs. Then, a neural network is designed to fuse these two modals for discriminating dementia patients from healthy controls. The network contains a cross-modal Transformer encoder for bimodal interaction, and a self-attention Transformer encoder for final classification. Experimental results demonstrate that the detection accuracy of the proposed method is 84.26%, which outperforms baseline methods and ablated models using single speech or eye-tracking input.

Index Terms— Dementia detection, bimodal fusion, speech, eye-tracking, bottleneck feature, Transformer

1. INTRODUCTION

Dementia is a neurodegenerative disease, which is characterized by cognitive disorder and memory decline, with Alzheimer's disease (AD) being the most common case [1]. It is estimated that 50 million people worldwide are living with dementia in 2019 and one new dementia patient will be diagnosed every three seconds [2]. The number of dementia patients is expected to 152 million by 2050. The annual medical cost of AD worldwide is about 1 trillion dollars, which is expected to almost double by 2030. At present, the diagnosis of dementia depends on cognitive tests, clinical manifestation, medical imaging, biomarker, etc., which are usually time consuming and expensive. Moreover, there is no effective treatment for dementia so far, so early detection and treatment are of great value in reducing the prevalence of dementia and delaying the progress of the disease.

Early patients with dementia suffer from language disorders [3]. Their performance on some language-based cognitive tasks, such as word-research ability, language fluency, and sentence repetition, are different from healthy people [4]. In order to detect dementia with speech, several speech datasets have been designed for dementia analysis [5–8]. Many researchers manually designed semantic features, auditory features, and other speech features for dementia detecting with machine learning algorithms [6, 9, 10]. Warnita et al. [11] extracted several INTERSPEECH feature sets from Pitt Corpus, which achieved the classification accuracy of 73.6% using a gated convolutional neural network (GCNN). Recently, Liu et al. [12] extracted speech BN features using an ASR model, which realized the automatic extraction of linguistics-related features, and then fed them into the neural network for AD detection. The accuracy of this method in DementiaBank corpus is 82.59 %.

In addition, the observation behaviors measured by eye-tracking data also show differences between dementia patients and healthy people as a result of difficulty in disengaging attention and impaired cognitive ability. In recent years, several studies have been conducted on dementia detection based on eye-tracking characteristics. Molitor et al. [13] summarized the effects of AD on eye saccade and pupil reflection, and found that patients showed random observational symptoms during visual search and eye movements appeared to be slower and less accurate. In the research of Oyama et al. [14], eye gaze position data was recorded when the subjects observed a series of specific movie clips, pictures and related tasks. The cognitive test scores, which were obtained by analyzing the above data, were positively correlated with the results of the Mini-Mental State Examination (MMSE). With the development of artificial intelligence, researchers tried to model eye movement with machine learning algorithms for AD detection. Lagun et al. [15] extracted eye movement features manually, and then support vector machines (SVM) and logistic regression (LR) were used for classification between AD and control (CTRL). Mengoudi et al. [16] proposed self-supervised representation learning to extract eye movement features, and high-level semantic coding was obtained by building convolutional neural networks.

Most previous studies on machine learning-based dementia detection focused on utilizing single modal data, e.g., speech or eye-tracking. However, multi-modal fusion has been proven to be an effective way to improving the performance of many tasks such as emotion recognition [17] and speech recognition [18]. To our best knowledge, this paper makes the first attempt of fusing speech and eye-tracking inputs for dementia detection. Inspired by the automatic speech recognition (ASR) based speech bottleneck (BN) features for AD detection [12], this paper designs a regional picture recognition (RPR) model as the eye-tracking feature extractor. The function of RPR is to recognize the class of the reginal picture that the subject is looking at (e.g, *boy*, *girl* or *chair*). Then, time-varying eye-tracking BN features are extracted from the last convolutional layer of the RPR model frame-by-frame and act as intermediate representations between the raw pixels and the class labels of regional pictures. Then, a neural network is built to fuse speech BN features and eye-tracking BN features for dementia detection, which contains a Transformer encoder with cross-modal attention (CMA)

This work was partially funded by the National Nature Science Foundation of China (Grant No. 62106246) and China Postdoctoral Science Foundation (Grant No. 2021M693101).

Table 1. The stasistics of the subjects in our dataset.

Group (number)	Gender (M/F)	Age mean(std)	Education mean(std)	MMSE mean(std)
CTRL(39)	18/21	59.6(14.3)	11.8(3.8)	27.1(2.0)
MCI(52)	16/36	62.4(11.5)	9.5(3.7)	24.3(3.2)
Dementia(11)	6/5	70.7(8.0)	11.9(2.8)	18.9(5.8)
ALL(102)	36/66	61.0(13.0)	10.5(3.8)	24.76(4.0)

for bimodal interaction and a self-attention Transformer encoder for further encoding and final classification.

Experiments were conducted on a picture description dataset with parallel speech and eye-tracking recordings. Our proposed method achieved a detection accuracy of 84.26%, which outperformed two baseline methods using single eye-tracking or speech input by 21.7% and 3.22% respectively. Some ablation experiments were also conducted to analyze the effectiveness of the blocks in our proposed method.

2. DATASET

The subjects in our dataset were recruited from the outpatient clinic of Shanghai Tongji Hospital. All subjects suffered from varying degrees of memory impairment and underwent a comprehensive physical examination and cognitive function test. According to guidelines of NIA-AA established in 2011 [19], all subjects were diagnosed into three categories, 11 patients with dementia, 52 patients with mild cognitive impairment (MCI), and 39 healthy people (CTRL). The statistics of the subjects' age, education, MMSE and MoCA are shown in Table 1. In order to make full use of the data, a two-category dementia detection task is employed in this paper, with one category containing both dementia and MCI subjects and another category containing only CTRL subjects.

The dataset consists of synchronized Chinese speech and eye-tracking recordings from the subjects during a picture description task. In the task, the subjects were asked to describe the picture of ”Cookie Theft” as much as possible. When encountering difficulties, they were provided specific tips by doctors. The subjects' descriptions on the picture and a few doctors' tips were recorded by a microphone with a sampling rate of 16 kHz. The total recording time of all subjects was 251 minutes. Speech recordings were manually annotated using the TextGrid format of Praat software, including the start-end positions, transcriptions, and speaker information of each utterance. Eye movement of each subject was recorded by Tobii eye tracker with a sampling rate of 125 Hz, which contained 78 attributes at each frame such as the position of sight, the state of eye movement, the duration of the state, the size of the pupil, etc.

3. METHODS

The task in the paper is to make a binary classification (Dementia/MCI or CTRL) based on the subject's speech and eye-tracking recordings. The overall flowchart of our proposed method is shown in Fig. 1, which contains several main blocks, including (1) bottleneck feature extractors, (2) a cross-modal Transformer encoder, and (3) a self-attention Transformer encoder.

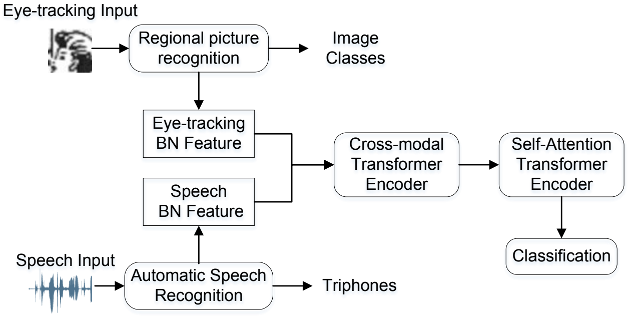
Eye-tracking Input
Regional picture recognition
Image Classes
Eye-tracking BN Feature
Speech BN Feature
Cross-modal Transformer Encoder
Self-Attention Transformer Encoder
Classification
Speech Input
Automatic Speech Recognition
Triphones

Fig. 1. The overall flowchart of our proposed method.

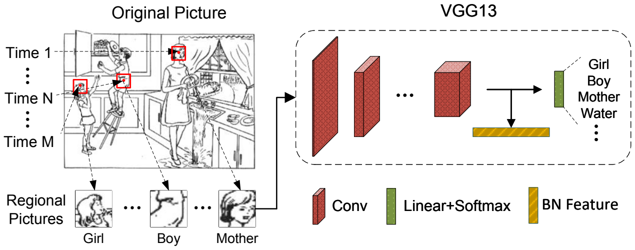

Fig. 2. The structure of the RPR model for extracting eye-tracking BN features. The original picture is adapted from Boston Diagonstic Aphasia Examination [20].

3.1. Content-based Bottleneck Feature Extraction

Eye movement and speech are synchronized bimodal information generated by a subject in the process of picture description, and we expect that content-based features can better bridge the gap of modal differences. Thus, before fusing these two modals, this paper trains deep learning models to extract BN features for each modal. BN features [21] were originally proposed by using a deep neural network (DNN) as a feature extractor for GMM-HMM speech recognition. And the DNN contained a hidden layer with fewer units than other layers, which was called the bottleneck layer. The activation values of this layer were extracted as bottleneck features, which provide intermediate representations between raw spectra and phonetic transcriptions for each frame.

Following the work of Liu et al. [12], this paper employs an LSTM-based speaker-independent ASR model to extract speech BN features. This model was trained using an internal dataset of iFLYTEK, which contained 3000 hours of Chinese recordings. The model inputs are MFCC features with a frame shift of 40ms, and the classification targets are clustered triphones as shown in Fig. 1. The outputs of its second linear layer are defined as speech BN features.

In order to describe what the subject is looking at according to his/her eye-tracking recordings, a regional picture recognition (RPR) model is built to extract eye-tracking BN features, as shown in Fig. 2. Here, the RPR model is a fine-tuned VGG13 model after comparing the performance of several classic CNNs. The original picture for data acquisition has a resolution of 865*622. We manually segmented the picture according to its contents and assigned 16-category class labels to all pixels. The classes include *boy*, *girl*, *mother*, *chair*, *closet*, *cake*, *curtain*, *tableware*, *floor*, *grass*, *house*, *road*, *table*, *tree*, *water*, and *others*. Then, 300 thousands 48*48-pixel regional pictures were generated by randomly cutting the original picture and each regional picture was assigned a class label according to the class of its central pixel. These regional pictures with class labels were used to finetune the VGG13 model. The activation vectors of the last convolutional layer were extracted as eye-tracking BN features.

We aligned the bimodal recordings based on timestamps and unified their sampling rates to 25 Hz. The dimensions of speech BN features and eye-tracking BN features at each frame are $d_S = 640$ and $d_{ET} = 512$ respectively. For the recordings of each subject, the sequences of two BN features were cyclically padded [22] (if the original sequence length is less than 1024) , truncated (if the original sequence length is between 1024 and 2048) or split (if the original sequence length is larger than 2048) to $T = 1024$, corresponding to the duration of 40.96 s for an instance. Furthermore, one-dimensional convolutions are applied to project bimodal BN features to the same dimension $d = 640$ and the output of this convolutions is position-encoded [23] to carry temporal information. Then, the aligned speech BN features $\boldsymbol{X}_S \in \mathbb{R}^{T\times d}$ and eye-tracking BN features $\boldsymbol{X}_{ET} \in \mathbb{R}^{T\times d}$ are sent into the following neural network blocks for classfication.

3.2. Cross-modal Transformer Encoder

In order to make better use of bimodal data, bimodal BN features $\boldsymbol{X}_S$ and $\boldsymbol{X}_{ET}$ are integrated in this block to generate new representatinons for classification. Inspired by multimodal emotion recognition [24], the cross-modal Transformer encoder is adopted here for bimodal interaction and its structure is shown in Fig. 3. The results of previous studies [12, 16] and our preliminary experiments have shown that using single speech modal can achieve better accuracy of dementia detection than using single eye-tracking modal. Thus, eye-tracking features are treated as auxiliary ones to enhance speech representations in this block.

As shown in Fig. 3, $\boldsymbol{X}^{[i]}_{S\_ET}$ are denotes the output of the i-th layer of the cross-modal Transformer encoder. Specially, $\boldsymbol{X}^{[0]}_{S\_ET} = \boldsymbol{X}_{ET}$ is used as input for the first layer. The core in the cross-modal Transformer encoder is the cross-modal attention mechanism (CMA). At the first layer, the output of CMA is calculated as

$$\begin{aligned} \boldsymbol{Y} &= \text{CMA}\left(\boldsymbol{X}_{ET}, \boldsymbol{X}_S\right), \\ &= \text{softmax}\left(\frac{\boldsymbol{X}_{ET}\boldsymbol{W}_{Q_{ET}}\boldsymbol{W}_{K_S}^{\top}\boldsymbol{X}_S^{\top}}{\sqrt{d_K}}\right)\boldsymbol{X}_S\boldsymbol{W}_{V_S}, \end{aligned} \tag{1}$$

where $\boldsymbol{W}_{K_S} \in \mathbb{R}^{d\times d_K}$ and $\boldsymbol{W}_{V_S} \in \mathbb{R}^{d\times d_V}$ are trainable matrices that map $\boldsymbol{X}_S$ into $\boldsymbol{K}_S$ and $\boldsymbol{V}_S$, i.e., the key and value vectors of CMA, respectively. $\boldsymbol{W}_{Q_{ET}} \in \mathbb{R}^{d\times d_K}$ is a trainable matrix that maps $\boldsymbol{X}_{ET}$ into $\boldsymbol{Q}_{ET}$, i.e., the query vectors of CMA.

Besides, $\text{softmax}(\cdot) \in \mathbb{R}^{T\times T}$ stands for a weight matrix, in which the (m, n) element is the attention weight calculated between the m-th query vector and the n-th key vector. The ouput of CMA is a matrix $\boldsymbol{Y} \in \mathbb{R}^{T_{ET}\times d_V}$, and its each row is the weighted sum of all rows in $\boldsymbol{V}_S$. In other layers, $\boldsymbol{X}^{[i-1]}_{S\_ET}$ replaces $\boldsymbol{X}_{ET}$ in Eq. (1).

The cross-modal Transformer encoder has 3 layers and the head number of multi-head attention is 5. We also built the cross-modal Transformer encoder by using speech features as auxiliary ones (i.e., reverse CMA) or using both features as auxiliary ones (i.e., dual CMA). The comparison results will be introduced in Section 4.3.

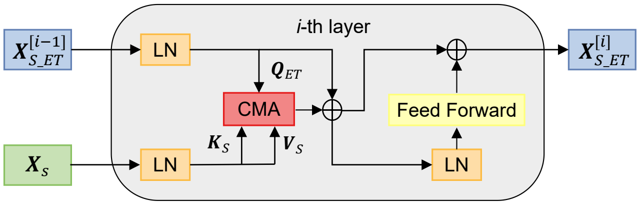

Fig. 3. The structure of the cross-modal Transformer encoder used in our method, where “LN” stands for “Layner Normalization”. In each layer, the speech features $\boldsymbol{X}_S$ are mapped to the key and value vectors of CMA, and the eye-tracking features $\boldsymbol{X}_{ET}$ (for the first layer) and the outputs of previous encoder layer $\boldsymbol{X}^{[i-1]}_{S\_ET}$ are mapped to the query vectors of CMA.

3.3. Self-attention Transformer Encoder and Classification

As shown in Fig. 1, the output of the cross-modal Transformer encoder is further processed by a self-attention Transformer encoder [23] before final decision. Its layer number is 2 and the head number or multi-head attention is 5. The last frame output of the self-attention Transformer encoder is classified into two categories through three fully connected layers. The first two layers contains 640 neurons with ReLU activation function and the output layer contains 2 neurons with softmax activation function. One dropout layer is inserted after each of the first two fully connected layers to reduce overfitting.

4. EXPERIMENTS

4.1. Model Implementation

Our models were trained to minimize the cross entropy loss function with the Adam optimizer in Pytorch. The weight decay of Adam was set as 5×10^{-4} to avoid overfitting. The batch size was set as 16 and the max epoch number was 150. We used a step-based learning rate scheduler. The initial learning rate was set at 0.001 and it dropped by 35% every 20 epochs. Early stop was also adopted to avoid overfitting.

Furthermore, two baseline models were built for comparison.

1. ET+SVM [15]: This method utilized single eye-tracking modal. It manually extracted nine eye-tracking features, including average gaze duration, average gaze amplitude, average gaze stability, average saccade duration, average saccade amplitude, average saccade speed, average pupil size, pupil size variance, and proportion of out-of-sight from the eye-tracking recordings of each subject. Traditional classifiers, including logistic regression, SVM, and naive Bayes were applied to build dementia detection models using above features. Finally, the SVM classifier was selected based on the detection performance.
2. S+CNN-LSTM [12]: This method utilized single speech modal. It adopted the same speech BN features as the ones used in our proposed method. The features were first processed by two-layer convolutional layers for local context modeling. The number of convolution kernels were 128 and 256 respectively, the size of the kernels was 3. Then, 2 BiLSTM layers were applied for global context modeling and each layer contained 48 units along each ditrection. Finally, attention pooling and fully connected layers were built for classification.

Table 2. Results of dementia detection using different methods (%).

Method	ACC	UAP	UAR
ET+SVM [15]	62.56	66.31	63.76
S+CNN-LSTM [12]	81.04	83.10	81.21
Our method	84.26	83.90	84.17

Table 3. Results of dementia detection in ablation studies (%).

Method	ACC	UAP	UAR
Our method	84.26	83.90	84.17
BN_S → LLDs	63.86	59.16	63.85
BN_ET → ET pictures	83.72	84.76	83.29
Single Eye-tracking	72.21	71.77	72.50
Single Speech	82.18	82.23	81.24
Reverse CMA	81.92	82.10	82.07
Dual CMA	84.36	83.77	84.21

4.2. Evaluation Metrics

Accuracy (ACC), unweighted average precision (UAP) and unweighted average recall (UAR) were employed as the evaluation metrics in our experiments. Considering the limited number of subjects in our dataset, 5-fold cross-validation was adopted to calculate above metrics in our experiments. The cross-validation was conducted 10 times with different random data segmentation. The means of the metrics among the 10 times of 5-fold cross-validation were reported.

4.3. Evaluation Results

The experimental results of the baselines and our proposed model are shown in Table 2. We can see that our method outperformed two baseline on all three metrics. There is a clear gap between the performance of ET+SVM and other two models. One possible reason is that the manually extracted eye-tracking features only contained temporally averaged physical descriptions of eye-tracking, and totally neglected the time-varying contents that the subject was looking at during the task. Comparing with S+CNN-LSTM, our method improved the dementia detection accuracy from 81.04% to 84.26%. This demonstrates the effectiveness of our proposed method based on fusing speech and eye-tracking inputs.

4.4. Ablation Studies

In order to verify the effectiveness of the extracted BN features, two ablated models were built by replacing speech BN features with 25 frame-level low-level descriptors (LLD) provided in the eGeMAPS acoustic feature set [25] (i.e., BN_S → LLDs) and replacing eye-tracking BN features with the raw pixels of regional picture sequences (i.e., BN_ET → ET pictures). The sampling rates of the LLDs and the regional picture sequences were the same as the BN features of two modals, and we adjusted the structure of convolution before cross-modal transformer encoder accordingly due to the change of input features. In the model of BN_ET → ET pictures, we used 2D-CNN for further eye-tracking feature extraction.

The dementia detection results of these two models were shown in the second and third rows of Table 3. Comparing with our proposed method, the performance of BN_S → LLDs degraded significantly. The advantages of speech BN features are that they contain linguistics-related information which may help the interaction with eye-tracking features and they reduce the diverse speaker characteristics and acoustic environments in raw acoustic features. Comparing the method of BN_ET → ET pictures with our proposed method, we can see that the RPR-based eye-tracking BN feature extraction contributed to achieve higher ACC and UAR. However, the benefit was much smaller than using speech BN features. One possible reason is that the raw regional pictures also contained content information which was simple in this task and may be extracted by following neural network calculations.

Furthermore, we adopted single speech BN features or single eye-tracking BN features as the input of the proposed model to check whether the cross-modal Transformer encoder for modal fusion was effective. For single-modal input, the model omitted the cross-modal Transformer encoder and directly sent bimodal BN features to the self-attention Transformer encoder with an increased number of layers. The results are shown in the fourth and fifth rows of Table 3. We can see that the accuracy of using single speech modal was 9.68% higher than that of using single eye-tracking modal. This supports our design in the cross-modal Transformer encoder that considers eye-tracking features as auxiliary ones to enhance speech features. Compared with the results of ET+SVM in Table 2, our method with single eye-tracking input improved the accuracy by 9.94%. This indicates the advantages of using content-related eye-tracking BN features instead of manually designed low-level descriptions and using Transformer-based sequence modeling insteand of simple SVM classifier. Our proposed method with model fusion improved the accuracies of using single speech or eye-tracking modal by 2.08% and 11.76% respectively. This confirms the effectiveness of the cross-modal interaction in our method.

Finally, we conducted experiments on the reverse CMA and dual CMA strategies mentioned in Section 3.3. In reverse CMA, speech was treated as the auxiliary modal, i.e., speech representations were mapped to queries and eye-tracking representations were mapped to keys and values. In dual CMA, two cross-modal Transformer encoder were built using either eye-tracking or speech as the auxiliary modal. Their outputs were processed by two self-attention Transformer encoder and were further concatenated for classification. The results of these two strategies were shown in the last two rows of Table 3. We can find that our method achieved 2.34% higher accuracy than reverse CMA and the performance of dual CMA was very close to our method. These results further confirms that our choice of using eye-tracking as the auxiliary modal in CMA is appropriate.

5. CONCLUSION

This paper have proposed a method of fusing speech and eye-tracking recordings of subjects in a picture description task for dementia detection. First, RPR and ASR models are built to extract content-related eye-tracking and speech BN features. Then, dementia detection is achieved by constructing a Transformer-based neural network with cross-modal interaction. Our proposed method achieved a detection accuracy of 84.26% in cross validation, which was better than the baseline and ablated models with singal modal input. To expand the data set, improve the multi-modal fusion algorithm and investigate the interpretability of the proposed model will be the tasks of our future work.

6. REFERENCES

[1] Maria Revi, "Alzheimer's disease therapeutic approaches," in *GeNeDis 2018*, pp. 105–116. Springer, 2020.

[2] Chris Lynch, "World Alzheimer report 2019: Attitudes to dementia, a global survey: Public health: Engaging people in adrd research," *Alzheimer's & Dementia*, vol. 16, pp. e038255, 2020.

[3] Mark W Bondi, Emily C Edmonds, and David P Salmon, "Alzheimer's disease: past, present, and future," *Journal of the International Neuropsychological Society*, vol. 23, no. 9-10, pp. 818–831, 2017.

[4] Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna 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, pp. 195, 2015.

[5] James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen 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.

[6] Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li, "Dementia detection by analyzing spontaneous mandarin speech," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 289–296.

[7] Ildikó Hoffmann, Dezso Nemeth, Cristina D Dye, Magdolna Pákáski, Tamás Irinyi, and János Kálmán, "Temporal parameters of spontaneous speech in Alzheimer's disease," *International journal of speech-language pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[8] Aharon Satt, Ron Hoory, Alexandra König, Pauline Aalten, and Philippe H Robert, "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[9] Maria K Wolters, Najoung Kim, Jung-Ho Kim, Sarah E MacPherson, and Jong C Park, "Prosodic and linguistic analysis of semantic fluency data: A window into speech production and cognition.," in *Interspeech*. San Francisco, CA, 2016, pp. 2085–2089.

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

[11] Tifani Warnita, Nakamasa Inoue, and Koichi Shinoda, "Detecting Alzheimer's disease using gated convolutional neural network from audio data," *arXiv preprint arXiv:1803.11344*, 2018.

[12] Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, and Yunxia Li, "Detecting Alzheimer's disease from speech using neural networks with bottleneck features and data augmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7323–7327.

[13] Robert J Molitor, Philip C Ko, and Brandon A Ally, "Eye movements in Alzheimer's disease," *Journal of Alzheimer's disease*, vol. 44, no. 1, pp. 1–12, 2015.

[14] Akane Oyama, Shuko Takeda, Yuki Ito, Tsuneo Nakajima, Yoichi Takami, Yasushi Takeya, Koichi Yamamoto, Ken Sugimoto, Hideo Shimizu, Munehisa Shimamura, et al., "Novel method for rapid assessment of cognitive impairment using high-performance eye-tracking technology," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[15] Dmitry Lagun, Cecelia Manzanares, Stuart M Zola, Elizabeth A Buffalo, and Eugene Agichtein, "Detecting cognitive impairment by eye movement analysis using automatic classification algorithms," *Journal of neuroscience methods*, vol. 201, no. 1, pp. 196–203, 2011.

[16] Kyriaki Mengoudi, Daniele Ravi, Keir XX Yong, Silvia Primativo, Ivanna M Pavisic, Emilie Brotherhood, Kirsty Lu, Jonathan M Schott, Sebastian J Crutch, and Daniel C Alexander, "Augmenting dementia cognitive assessment with instruction-less eye-tracking tests," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3066–3075, 2020.

[17] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[18] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[19] Guy M McKhann, David S Knopman, Howard Chertkow, and others Hyman, "The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[20] Harold Goodglass, Edith Kaplan, and Barbara Barresi, *Boston diagnostic aphasia examination record booklet*, Lipppincott Williams & Wilkins, 2001.

[21] Frantisek Grézl, Martin Karafiát, Stanislav Kontár, and Jan Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 4, pp. IV–757.

[22] Jian Huang, Ya Li, Jianhua Tao, Zhen Lian, et al., "Speech emotion recognition from variable-length inputs with triplet loss function.," in *Interspeech*, 2018, pp. 3673–3677.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[24] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2019, vol. 2019, p. 6558.

[25] Florian Eyben, Klaus R Scherer, Björn W Schuller, 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.