END-TO-END ASR-ENHANCED NEURAL NETWORK FOR ALZHEIMER'S DISEASE DIAGNOSIS

Jiancheng Gui, Yikai Li, Kai Chen, Joanna Siebert, Qingcai Chen✉

Harbin Institute of Technology (Shenzhen), China
gjc.hit@qq.com, 190110419@stu.hit.edu.cn, ckai.hit@gmail.com
joannasiebert@yahoo.com, qingcai.chen@hit.edu.cn

ABSTRACT

This paper presents an approach to Alzheimer's disease (AD) diagnosis from spontaneous speech using an end-to-end ASR-enhanced neural network. Under the condition that only audio data are provided and accurate transcripts are unavailable, this paper proposes a system that can analyze utterances to differentiate between AD patients, healthy controls, and individuals with mild cognitive impairment. The ASR-enhanced model comprises automatic speech recognition (ASR) with an encoder-decoder structure and the encoder followed by an AD classification network. The encoder takes a Mel spectrogram as input and transforms it into high-level acoustic features that correlate with AD. The classification network then maps intermediate acoustic features to three categories. In the training phase, the AD classification and speech recognition tasks are optimized simultaneously. Experimental results obtained from an AD recognition dataset of Chinese spontaneous speech[1] illustrate the effectiveness of integrating ASR into AD diagnosis in an end-to-end manner. Further, our model has low dependency on accurate ASR transcripts. This work achieved accuracy scores of 89.1% and 82.6% for long and short utterance tracks, respectively.

***Index Terms*—** AD diagnosis, end-to-end, ASR, classification network, AD recognition challenge

1. INTRODUCTION

Alzheimer's disease (AD) is an irreversible chronic neurodegenerative disorder. Its clinical manifestations are progressive memory function loss, slow movement, slurred speech, and cognitive decline. Mild cognitive impairment (MCI) is one of the earliest indicators of AD.

As the average age of the population increases, an increasing number of older people suffer from AD [1]. However, early diagnosis and treatment of AD effectively slow down or arrest the development of the disease. Thus, early diagnosis of AD presents not only immense research value but also great clinical value. AD diagnosis is usually modeled as a classification task that the utterance is mapped to one of three categories: AD, MCI, and healthy controls (HC). Lately, an increasing number of studies have been conducted to facilitate early AD diagnosis based on spontaneous speech [1].

Currently, AD diagnosis from spontaneous speech can be roughly divided into two main categories. One is a linguistic method that attempts to analyze various linguistic features from speech transcripts [2]. Li et al. [3] found that TF-IDF vector and BERT embeddings can detect AD with high accuracy. Yuan et al. [4] applied pause-encoded transcripts to retrain language models, which outperformed text-only language models. Koo et al. [5] used pretrained language models to extract textual features and adopted three kinds of handcrafted textual features as additional input: such as psycholinguistic, repetitiveness, and lexical complexity features.

The other is a phonetic method that attempts to analyze various acoustic features from audio. Szatloczki et al. [6] found that temporal features such as the number and duration of pauses, phonation time, phonation-to-time ratio, and articulation rate correlate with the severity of AD. Meilán et al. [7] used temporal and acoustic features from speech and detected AD with high accuracy. Amit et al. [8] employed a log-mel spectrogram and MFCC for three different deep neural networks for AD recognition. In addition, x-vectors [3, 9], i-vectors [10], bag of audio words (BOAW) [11], and features extracted by openSMILE [12] followed by diverse classifier backends have been used, with varying performance levels.

Moreover, many methods combine linguistic and phonetic features [5, 10, 13]. Studies have shown that text feature-based methods outperform acoustic feature-based methods [3, 5, 10, 11]. However, manual transcripts are rarely available, and there are no specific automatic speech recognition (ASR) systems for AD. This results in many substitution errors in ASR transcripts owing to accents and the speaking characteristics of older people. Thus, using trustless transcribed text leads to poor performance [14]. The performance of text feature-based methods needs to be guaranteed with accurately transcribed text, while audio data have advantages such as easy collection and rich semantic and emotional information. Thus, this research mainly aimed to investigate how to better

[1] AD recognition challenge: https://github.com/THUsatlab/AD2021

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

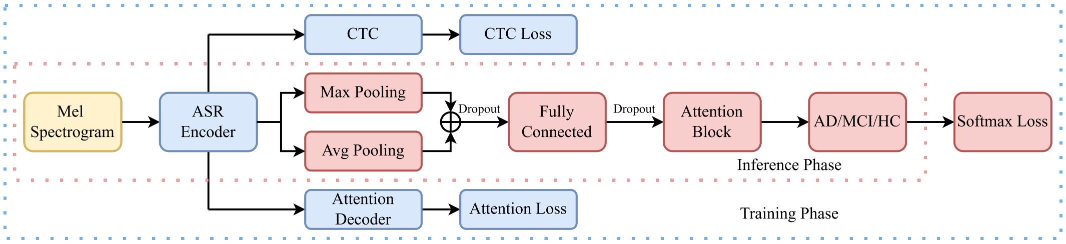


Fig. 1. The proposed ASR-enhanced AD diagnosis model.

extract the acoustic characteristics that correlate with AD.

To solve the above problems, a useful solution is to use the bottleneck (BN) features extracted by a pretrained ASR system [15] in which their classification network does not interact with the ASR module because of ASR errors. Aparna at al. [14] found that AD detection is most influenced by ASR word deletion errors. Tóth at al. [16] posited that ASR can be applied to perform phonetic-level segmentation and annotation to characterize the speaker's articulatory and phonetic traits. Studying audio and its ASR transcripts, we noticed that older people with cognitive impairment showed more substitution errors in the transcripts. Inspired by [15] and [16], we did not directly use ASR transcripts to extract linguistic features but integrated encoder-decoder-structured ASR with an AD classification network, hoping that this end-to-end structure could help the ASR-enhanced network automatically learn high-level representations oriented toward the AD classification task instead of extracting BN features ahead of time [15]. Moreover, multitask training in both ASR and classification can help the network learn better.

The main contributions of this paper are as follows. First, we propose integrating ASR into an AD diagnosis network to automatically learn the acoustic features that characterize AD patients in an end-to-end manner. Second, our system adopts a multitask learning mechanism to jointly optimize the ASR and AD classification tasks, which can help alleviate overfitting in AD detection in small datasets and improve the robustness of the AD diagnosis system.

2. METHODS

In this section, we first describe the overall system framework for AD diagnosis shown in Figure 2. We then introduce the methods used in the AD diagnosis model.

2.1. Overview of the AD Diagnosis System

The overall process of AD diagnosis consists of data preparation, training, and inference, as shown in Figure 2. The corpus comes from the AD recognition challenge, which includes audio data, collected from cognitive tests. To realize automatic AD diagnosis, we used speech transcripts generated by a commercial iFLYTEK ASR system as the training input. The training and inference processes of the diagnosis system are described in detail in Figure 1. In the training phase, a low-level acoustic feature Mel spectrogram is inputted into the ASR-enhanced network, which consists of the encoder module of an encoder-decoder ASR, a subsequent classification network, a connectionist temporal classification (CTC) module, and an attention decoder. By integrating ASR into the AD classification network, the ASR encoder and subsequent classification network can automatically learn the AD diagnosis task-oriented high-level features in an end-to-end way. To alleviate the influence of overfitting and utilize the informative notional words and frequent occurrence of the function words in transcripts, training in both the classification and ASR tasks is applied to learn robust high-level acoustic features. In the inference phase, only audio data are required for the diagnosis system, which then outputs the diagnosis results from one of the following three categories: AD, MCI, or HC.

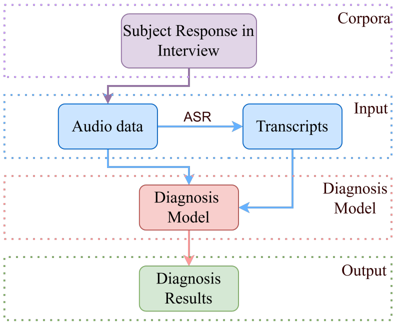


Fig. 2. Overall procedure for AD diagnosis.

2.2. Diagnosis Model

In this section, we introduce the methods used in ASR-enhanced AD diagnosis network in Figure 1, which consists of several parts, including an ASR encoder, an AD classification network with an attention block, and ASR auxiliary training for the classification task.

2.2.1. ASR Encoder

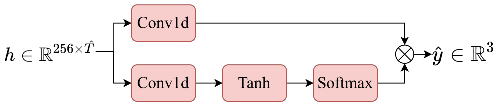
Although using ASR to extract high-level acoustic features for AD diagnosis is popular for speech-related tasks [15, 16, 17], we did not build our AD diagnostic system directly on these features for two reasons. First, AD causes changes in speech patterns, such as fluency, number and duration of pauses, repetition, and lower articulatory rate. Modeling an AD classification directly on acoustic features extracted from ASR cannot take full advantage of the rich semantic information contained in utterances, particularly when the amount of training data is limited. Second, although these acoustic features are high-level, they are not task-oriented; for example, the BN features extracted from ASR usually contain phoneme-level segmentations and annotations, which are not immediate biomarkers of AD.

To address the above problems, we integrated ASR into the AD classification network. With the help of the ASR module, the ASR-enhanced network can not only learn and exploit AD-related biomarker features, such as the articulatory rate and length of breaks, in an end-to-end manner but also utilize deictic words, such as "this," "that," and "those", and indefinite pronouns, such as "some," "few," "every," that appear in transcripts.

We use WeNet[18], a pretrained ASR model, detailed structure about encoder and decoder can refer to $multi\_cn$ example.[2] We denote the input of the ASR encoder as $X \in \mathbb{R}^{80\times T}$, and then we have output $X_{enc} \in \mathbb{R}^{\hat{T}\times D}$, where $\hat{T} = T/4$ is the spatial size along the time axis, and $D = 256$ is the embedding size for each acoustic feature frame. The acoustic features are then sent to the AD classification network and two decoder branches.

2.2.2. *AD Classification Network with an Attention Block*

To map high-level acoustic features X_{enc} to specified categories, we use the classification network depicted in Figure 1. The network contains a spatial aggregation module, a fully connected module, and an attention block. To aggregate local features, average pooling and max pooling are used in two parallel branches. The kernel size is 3, and stride is set to 1 to retain the time axis resolution. The aggregated features are then added together. The dropout unit is subsequently applied at a 0.5 keep rate. To increase nonlinearity, a fully connected layer that preserves dimensionality with rectified linear unit activation is adopted, and another similar dropout unit follows it. Finally, considering that high-level acoustic features contribute differently to determining the decision boundary for AD classification, the attention mechanism is used to obtain a weighted sum of each frame's confidence. As shown in Figure 3, there are two Conv1d modules, and output dim is 3, which are used to compute the attention weights and frame-wise prediction, respectively. To this end, we have input $h \in \mathbb{R}^{256\times\hat{T}}$ as the module input, through two Conv1d branches; then, we have $h_{att} \in \mathbb{R}^{3\times\hat{T}}$ and $h_{cls} \in \mathbb{R}^{3\times\hat{T}}$. The mathematical description is as follows:



$h \in \mathbb{R}^{256\times\hat{T}}$
Conv1d
Conv1d
Tanh
Softmax
$\hat{y} \in \mathbb{R}^3$

Fig. 3. CNN-based attention block.

$$h_{att} = W_a * h \tag{1}$$
$$h_{cls} = W_c * h \tag{2}$$
$$a = softmax(tanh(h_{att})) \tag{3}$$
$$\hat{y} = h_{cls} \odot a \tag{4}$$

The left branch calculates the attention-weighted matrix a, the right branch calculates the frame-wise prediction h_{cls}, and W_a and W_c are trainable weights in Conv1d. Then, the prediction $\hat{y}$ for utterance is obtained from the dot product and weighted sum.

2.2.3. *Using ASR for Auxiliary Training*

To fully utilize the ASR model, multitask training for classification and ASR is adopted. Assumimg that y is the ground truth label for utterance, the classification task for AD diagnosis can be constrained by:

$$L_{cls} = CE(\hat{y}, y) \tag{5}$$

where CE refers to the cross entropy function. Due to the small size of the dataset for AD diagnosis, if the acoustic model is directly used to model the classification task, it will result in overfitting. Thus, we need to use the pretrained ASR model to perform the auxiliary training classification task. The loss function for the ASR task is calculated as:

$$L_{asr} = \alpha * l_{ctc} + (1-\alpha) * l_{att} \tag{6}$$

where l_{ctc} is the CTC decoding loss, l_{att} is the attention decoding loss, and the hyperparameter α balances the importance of the two ASR loss functions. Thus, we obtain the final multitask training loss:

$$L = L_{cls} + \beta * L_{asr} \tag{7}$$

where β is the learning weight for ASR auxiliary training for the classification task.

3. EXPERIMENTS

3.1. Datasets

The competition provided 280 Chinese speech samples for training including 26 AD subjects aged between 62 and 82 years, 52 MCI subjects aged between 54 and 82 years, 44 HC subjects aged between 58 and 80 years, and 79, 93, and 108 samples for AD, MCI, and HC, respectively. The samples showed obvious stitching traces, and their lengths ranged

[2] WeNet: https://github.com/wenet-e2e/wenet

Table 1. Experimental results of all metrics for AD diagnosis based 6-s utterances using different models(%).

Model	Feature	Precision	Recall	F_1	Accuracy
ERNIE	Words	57.5	56.6	55.8	58.1
SVM	Manual	74.8	73.7	73.3	75.4
ResNet18	Mel	74.4	73.3	72.5	75.0
ResNet18	Spectrum	74.9	73.7	73.5	75.0
ResNet18	MFCC	69.3	67.4	66.0	69.4
ResNet18-LSTM	Mel	75.8	74.6	74.0	76.3
CNN-BiLSTM	Mel	73.0	72.5	72.4	73.7
CNN-BiLSTM	BN	78.8	78.2	78.0	79.4
Ours	**Mel**	**82.5**	**82.0**	**82.1**	**82.6**

Table 2. Experimental results of accuracy scores for AD diagnosis based on 60-s utterances using different models(%).

Model	Feature	Accuracy
ERNIE	Words	61.0
SVM	Manual	79.8
ResNet18	Mel	79.8
ResNet18-LSTM	Mel	81.5
CNN-BiLSTM	Mel	79.4
CNN-BiLSTM	BN	85.7
Ours	**Mel**	**89.1**

from 30 to 60 s. These samples were obtained from interviews with the subjects, which involved picture descriptions, fluency tests, and free conversation tasks. Most samples exhibited signs of an accent. The test set contained 119 long (60-s) and 1153 short (6-s) utterances.

3.2. Training and Evaluation Settings

The proposed system was used to perform both long and short utterances classification tasks. In the training stage, 280 long samples were divided into 6-s segments to obtain 2508 short clips, and then transcripts were retrieved using the iFLYTEK ASR API. The short utterance samples were divided into train and validation sets in a 4:1 ratio, the learning rate was set to 0.002, and the hyperparameters α and β were set to 0.3 and 0.1, respectively.

To evaluate the performance of our proposed system, we conducted several experiments using previous methods. First, we finetuned ERNIE [19] with no pause-encoded transcripts. Second, we manually extracted 1582 dim acoustic features using openSMILE [20] and used them to train an SVM classifier. Third, we evaluated some neural network methods used in AD classification tasks that can automatically learn task-oriented features, such as ResNet18 and ResNet18-LSTM [8] and CNN-BiLSTM [15], with a Mel spectrogram and BN features extracted by our ASR encoder. To evaluate the model's performance on long utterances, we divided 60-s utterances into multiple 6-s clips and obtained the predictions for the whole utterances by a majority vote.

Table 3. Ablation study of our model(%).

Model	Accuracy
Ours	**82.6**
- End-to-end manner	63.8
- ASR auxiliary training	81.9
- Attention block	81.1

3.3. Comparison and Ablation Studies

We compared our system with previous works using this dataset (Tables 1 and 2). Directly using the text feature-based methods showed poor performance, illustrating the unreliability of ASR transcripts. Acoustic features, such as Mel spectrograms, yielded better performance when finetuned on ResNet18. Adding an LSTM layer after ResNet18 resulted in further improvement. Moreover, ASR-based BN features resulted in even higher accuracy. To our knowledge, the best reported accuracy on this dataset is 89.9% for long utterances and 85.6% for short utterances. Our model achieved the second-highest accuracy.

To analyze the importance of different parts of the system quantitatively, we trained the models using several settings. First, we did not use end-to-end manners to train the model instead of using pre-extracted BN features by ASR encoder as the input of classfication network as in [15]. Second, we optimized the model without ASR parameter initialization and trained it with classification loss. Third, we replaced the attention block with a Conv1d. As shown in Table 3, we proved the advantage of integrating ASR into the AD classification task in an end-to-end manner with multitask training for both ASR and AD classification. Our attention block also achieved a slight improvement.

4. CONCLUSION

In this paper, we propose an ASR-enhanced neural network for AD diagnosis for which accurate manual transcripts are unnecessary. Integrating ASR into the classification network can help automatically learn in both the ASR and AD classification tasks, which helps our model learn discriminative acoustic representations. In future work, it is important to investigate the reasons for which the ASR-enhanced model can tolerate articulatory errors. Moreover, an elaborately designed classification network could be devised in an end-to-end manner to improve performance.

5. ACKNOWLEDGEMENTS

This work is supported by the Natural Science Foundation of China (Grant No. 61872113), Strategic Emerging Industry Development Special Funds of Shenzhen (No. XMHT20190108009), and Fundamental Research Fund of Shenzhen (No. JCYJ20190806112210067).

6. REFERENCES

[1] María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, et al., "Alzheimer's disease and automatic speech analysis: a review," *Expert systems with applications*, vol. 150, pp. 113213, 2020.

[2] Romola S Bucks, Sameer Singh, Joanne M Cuerden, et al., "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.

[3] Jinchao Li, Jianwei Yu, Zi Ye, et al., "A comparative study of acoustic and linguistic features classification for alzheimer's disease detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6423–6427.

[4] Jiahong Yuan, Xingyu Cai, and Kenneth Church, "Pause-encoded language models for recognition of alzheimer's disease and emotion," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7293–7297.

[5] Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, et al., "Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition," *arXiv preprint arXiv:2009.04070*, 2020.

[6] Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, et al., "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.

[7] Juan José G Meilán, Francisco Martínez-Sánchez, Juan Carro, et al., "Speech in alzheimer's disease: can temporal and acoustic parameters discriminate dementia?," *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.

[8] Amit Meghanani, CS Anoop, and AG Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 670–677.

[9] Raghavendra Pappagari, Jaejin Cho, Laureano Moro-Velazquez, et al., "Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity.," in *INTERSPEECH*, 2020, pp. 2177–2181.

[10] Anna Pompili, Thomas Rolland, and Alberto Abad, "The inesc-id multi-modal system for the adress 2020 challenge," *arXiv preprint arXiv:2005.14646*, 2020.

[11] Nicholas Cummins, Yilin Pan, Ren, 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] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Openear—introducing the munich open-source emotion and affect recognition toolkit," in *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 2009, pp. 1–6.

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

[14] Aparna Balagopalan, Ksenia Shkaruta, and Jekaterina Novikova, "Impact of ASR on alzheimer's disease detection: All errors are equal, but deletions are more equal than others," in *Proceedings of the Sixth Workshop on Noisy User-generated Text, W-NUT@EMNLP 2020 Online, November 19, 2020*, 2020, pp. 159–164.

[15] Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, et al., "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.

[16] László Tóth, Gábor Gosztolya, Veronika Vincze, et al., "Automatic detection of mild cognitive impairment from spontaneous speech using asr," ISCA, 2015.

[17] Noé Tits, Kevin El Haddad, and Thierry Dutoit, "Asr-based features for emotion recognition: A transfer learning approach," *arXiv preprint arXiv:1805.09197*, 2018.

[18] Binbin Zhang, Di Wu, Zhuoyuan Yao, et al., "Unified streaming and non-streaming two-pass end-to-end model for speech recognition," *arXiv preprint arXiv:2012.05481*, 2020.

[19] Yu Sun, Shuohuan Wang, Yukun Li, et al., "Ernie 2.0: A continual pre-training framework for language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 8968–8975.

[20] Björn Schuller, Stefan Steidl, Anton Batliner, et al., "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.