ALZHEIMER'S DETECTION FROM SPEECH USING SUPERVISED EMBEDDINGS WITH TEMPORAL ATTENTION

Jamshaid Iqbal[1], Sadaf Rehmat[1], Jiyun Li[1], Hira Sabir[2], Arslan Malik[3] and Mehtab Zafar[4]

[1]School of Information and Intelligent Science, Donghua University, Shanghai, China
[2] Department of Information Technology, University of Southern Punjab, Multan,Pakistan
[3]School of TransportationEngineering, ChangAnUniversity,Xian, China
[4]School of Computer Science and Technology, Jiangsu University of Science and Technology, Zhenjiang, China

ABSTRACT

Speech-based assessment of cognitive decline offers a low-burden and non-invasive alternative to resource-intensive diagnostic procedures. However, many existing systems rely on transcripts, multimodal fusion, or fixed acoustic features, which increase pipeline complexity and can limit practical deployment. This paper presents a transcript-free speech-only framework for three-way classification of Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy controls (HC) from spontaneous speech. The proposed method combines supervised fine-tuning of a pretrained Wav2Vec2.0 encoder with lightweight auxiliary acoustic cues and temporal attention for recording-level prediction. Experiments on the NCMMSC-AD 2021 benchmark are conducted under a strict patient-disjoint hold-out protocol in both short-speech and long-speech settings. The proposed system achieves 85.19% accuracy and 85.69% macro-F1 in the short-speech setting, and 88.04% accuracy and 87.57% macro-F1 in the long-speech setting. These results show that task-adapted acoustic representation learning and temporal aggregation can provide strong transcript-free performance for cognitive-status classification.

Keywords

Alzheimer's detection, Supervised Embeddings, Speech, Wav2vec2, Mild cognitive impairment

1. INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia and a growing global health challenge [1]. Because disease-modifying treatment remains limited and clinical decline is often gradual, early identification is especially important. Recognizing impairment at the mild cognitive impairment (MCI) stage can support earlier intervention, closer monitoring, and better care planning [2]. However, widely used diagnostic tools, including neuroimaging and cerebrospinal fluid biomarkers, are costly, time-consuming, and in some cases invasive, which restricts their use in routine screening and repeated follow-up [3]. These limitations have motivated increasing interest in low-burden, scalable, and repeatable assessment strategies.

Speech is one of the most promising candidates for such assessment [4]. Cognitive decline can affect multiple aspects of spoken communication, including fluency, pausing behavior, prosody,

articulation, and broader discourse organization [5][6]. Unlike many conventional biomarkers, speech can be collected non-invasively with common recording devices and can therefore support frequent monitoring in real-world or remote settings. This combination of clinical relevance, accessibility, and low acquisition cost has made speech-based analysis an attractive direction for automatic AD screening and longitudinal cognitive assessment[7].

A large body of prior work has explored machine learning for AD detection from speech. Many high-performing systems incorporate transcripts because lexical, syntactic, and semantic abnormalities are informative markers of cognitive decline[8]. Nevertheless, transcript-dependent pipelines require either manual annotation or automatic speech recognition (ASR), which increases cost, adds system complexity, and introduces an additional source of error. These constraints are particularly important in practical and low-resource settings, where transcription quality may be inconsistent and end-to-end deployment must remain simple. As a result, transcript-free speech-only diagnosis has become an increasingly important research direction. Recent advances in pretrained speech representation learning, especially Wav2Vec2.0, have further strengthened this direction by enabling rich acoustic modeling directly from raw waveforms [9].

These challenges are well illustrated by the NCMMSC-AD 2021 benchmark, which has become an important testbed for three-way classification of AD, MCI, and healthy controls (HC). Existing speech based studies on this dataset have explored several strategies, including raw-speech modeling with adapted pretrained ASR backbones [10], attention-augmented CNNs on acoustic representations , and more recent self-supervised or pretrained encoder-based frameworks [11, 12]. Although these studies demonstrate the promise of speech-only diagnosis, two limitations remain. First, the dataset is still relatively limited in size, which makes robust representation learning difficult and increases the risk of overfitting. Second, diagnostically informative evidence is often sparse and unevenly distributed within long recordings, so performance depends not only on the quality of the learned acoustic representations, but also on how informative regions within a recording are emphasized and aggregated over time.

Motivated by these observations, we propose a transcript-free speech-only framework for three-way cognitive-status classification from spontaneous speech. The method combines supervised fine-tuning of a pretrained Wav2Vec2.0 encoder with lightweight auxiliary speech-derived cues and an attention-based temporal classifier for recording-level prediction. Unlike approaches that rely on fixed acoustic features, simple segment voting, or transcript-dependent processing, the proposed framework is designed to strengthen both acoustic representation learning and temporal aggregation within a unimodal pipeline.

Evaluated on NCMMSC-AD 2021 under a strict patient-disjoint hold-out protocol, the proposed system is assessed in both short-speech and long-speech settings. The paper makes three contributions. First, it presents a speech-only framework that combines supervised pretrained acoustic representation learning with temporal attention for three-way AD/MCI/HC classification. Second, it studies this framework under patient-level evaluation in both short- and long-speech conditions, clarifying the effect of recording duration on diagnostic performance. Third, it shows that carefully optimized unimodal acoustic modeling can provide strong transcript-free performance without relying on explicit linguistic features or multimodal fusion.

2. RELATED WORK

The NCMMSC-AD 2021 benchmark has become an important testbed for speech-based three-way classification of Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy

controls (HC). Early speech-only systems on this dataset mainly relied on acoustic features and relatively lightweight downstream classifiers. The official challenge baseline reported 79.8% accuracy on the long-speech track and 74.0% on the short-speech track. Building on this benchmark, Qin et al. [10] proposed a raw-speech approach based on an adapted pretrained ASR model and reported 83.2% and 78.0% accuracy on the long- and short-speech tracks

More recent studies have shifted from handcrafted acoustic representations toward pretrained and self-supervised speech encoders. Yang et al. [12] introduced an augmented adversarial self-supervised learning framework and reported a macro-F1 of 83.73 on NCMMSC-AD 2021, while Zhang et al. [13] investigated Wav2Vec2/XLSR-53 representations with downstream GRU and attention-based classifiers, reporting test accuracy up to 85.45 under their experimental split. Taken together, these studies suggest that speech-only diagnosis depends not only on the quality of learned acoustic representations, but also on how informative regions are emphasized and aggregated across time within each recording.

Building on this line of work, the present study adopts a speech-only framework that combines supervised fine-tuning of a pretrained Wav2Vec2.0 encoder with lightweight auxiliary cues and an attention-based temporal classifier. Compared with earlier systems, the proposed method is designed to improve task-adapted acoustic representation learning within a simple transcript-free pipeline.

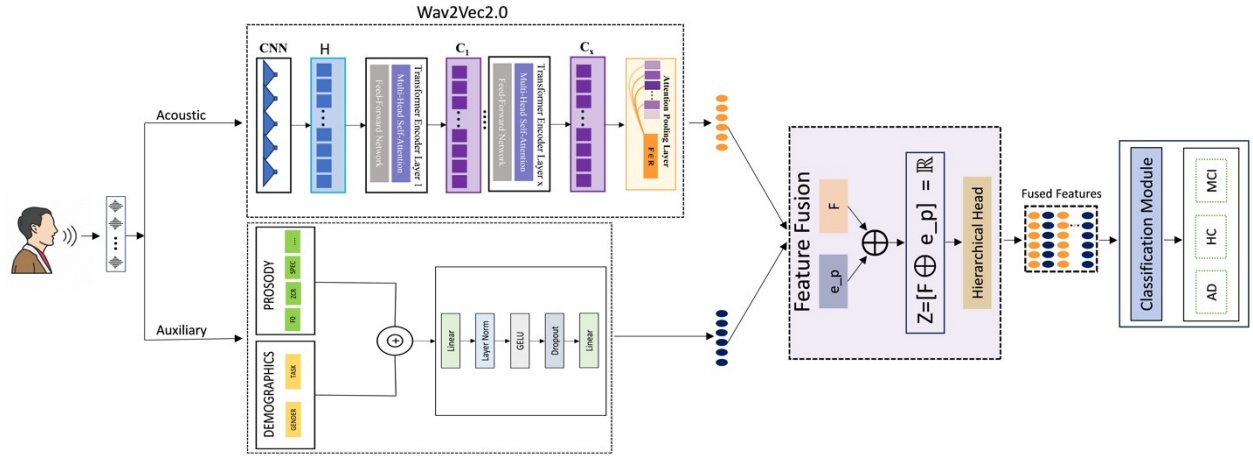

Figure 1: Overall Workflow of the Model.

3. METHODOLOGY

3.1. Dataset and Preprocessing

Experiments were conducted on the NCMMSC-AD 2021 dataset, a Chinese speech benchmark for three-way classification of Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy controls (HC) [14]. The dataset includes both long-speech and short-speech conditions. Long recordings are approximately 30–60 s in duration, whereas short-speech samples are 10 s segments extracted from the long recordings. The corpus contains 123 unique subjects with both single-session and multi-session recordings. To preserve subject integrity, evaluation was performed at the patient level, with all sessions from the same subject kept in the same split. A dedicated hold-out test set of 44 subjects was used for final evaluation, including 9 AD, 19 MCI, and 16 HC subjects, while the remaining 79 subjects were used for model development.

All audio data underwent a standardized preprocessing pipeline. Each recording was converted to mono and resampled to 16 kHz. For model input standardization, long recordings were segmented into non-overlapping 10-second windows, if the final part of a recording was shorter than 10 seconds, it was padded with silence to reach the required input length for Wav2Vec2.0. All inputs were subsequently normalized using z-score normalization.

3.2. Acoustic Features

Wav2Vec2.0 [9] is used as the acoustic backbone of the proposed framework because it provides strong contextual speech representations directly from raw audio and can be adapted effectively to downstream clinical classification tasks. In the present study, the overall acoustic branch is illustrated in Fig. 1. Each recording is first standardized and divided into fixed-length 10-second speech segments, which serve as the basic input units for representation learning and subsequent temporal modeling.

Each segment is then processed by the pretrained Wav2Vec2.0 encoder. The low-level convolutional feature extractor is kept frozen, while the higher contextual layers are fine-tuned on the target task. This design preserves the robustness of the pretrained acoustic front end while allowing the contextual representations to adapt to cognitively relevant speech patterns in spontaneous clinical speech. To improve robustness under limited training data, waveform-level augmentation is applied during fine-tuning, including controlled noise injection, time masking, and mild speed perturbation.

After contextual encoding, the latent speech sequence is not reduced by simple average pooling. Instead, it is first refined with a multi-head self-attention module so that each time step can incorporate information from other parts of the same segment. The refined sequence is then combined with the original contextual representation through residual fusion and normalization. Based on this refined representation, temporal importance weights are computed across the sequence, and the final acoustic embedding is obtained as a weighted summary of the latent time steps.

3.3. Auxiliary Features

To complement the deep contextual acoustic embedding, a lightweight auxiliary branch was introduced to capture compact prosodic and demographic cues at the segment level. For each speech segment, five low-level descriptors were extracted directly from the waveform: signal energy, zero-crossing rate, spectral centroid, spectral roll-off, and spectral flux. Together, these features provide a compact summary of voicing activity, temporal variation, and coarse spectral behavior that may still carry useful information for cognitive-status discrimination.

In addition to these prosodic descriptors, gender information was incorporated through a learned embedding. The prosodic feature vector and the gender embedding were concatenated to form the auxiliary input for each segment. This combined input was then passed through a lightweight feed-forward projection block composed of a linear transformation, layer normalization, GELU activation, dropout, and a second linear layer to obtain the final auxiliary representation.

The resulting auxiliary vector provides a low-cost complementary description of each speech segment and is fused with the deep acoustic embedding in the next stage. In this way, the auxiliary branch enriches the segment-level representation with compact handcrafted and demographic cues while keeping the overall framework lightweight and fully speech-based.

3.4. Feature Fusion

After acoustic and auxiliary processing, the two branches are fused by concatenation,

$$z_i = [a_i; q_i], z_i \in \mathbb{R}^{832} \quad (1)$$

The fused vector $z_i$integrates contextual acoustic information from Wav2Vec2.0 with compact prosodic and demographic cues.

To further enhance the discriminative quality of the fused representation, hierarchical supervision is introduced through two auxiliary binary heads. The first head captures the coarse distinction between healthy controls and cognitively impaired subjects,

$$\ell_i^{(h)} = f_h(z_i), \ell_i^{(h)} \in \mathbb{R}^2 \quad (2)$$

whereas the second head capture the finer distinction between AD and MCI,

$$\ell_i^{(s)} = f_s(z_i), \ell_i^{(s)} \in \mathbb{R}^2 \quad (3)$$

Rather than using these auxiliary outputs as isolated side predictions, their logits are concatenated back with the fused representation and passed to the main diagnostic head,

$$\tilde{z}_i = [z_i; \ell_i^{(h)}; \ell_i^{(s)}], \tilde{z}_i \in \mathbb{R}^{836} \quad (4)$$

$$\ell_i^{(m)} = f_m(\tilde{z}_i), \ell_i^{(m)} \in \mathbb{R}^3 \quad (5)$$

In this way, the coarse healthy-versus-impaired cue and the finer AD-versus-MCI cue are reinjected into the fused representation, making the learned features more diagnostically informative. This design encourages the representation to encode both coarse and fine-grained clinical structure instead of relying only on a single supervisory signal.

The overall training objective is defined as

$$\mathcal{L} = 0.6\,\mathcal{L}_{\text{main}} + 0.2\,\mathcal{L}_{\text{healthy}} + 0.2\,\mathbf{1}[y_i \neq HC]\,\mathcal{L}_{\text{severity}} \quad (6)$$

where $L_{\text{main}}$is a class-weighted three-class cross-entropy loss, and $L_{\text{healthy}}$and$L_{\text{severity}}$are two-class cross-entropy losses for the auxiliary tasks. Through this hierarchical objective, the model is encouraged to learn fused features that better reflect the underlying clinical progression from healthy control to MCI and AD.

3.5. Classification

Following feature extraction, each subject is represented as a temporal sequence of fused feature vectors. Since speech duration varies across subjects and sessions, a flexible sequential representation is adopted instead of a rigid fixed-length input. Consecutive speech segments are arranged into variable-length sequences, with shorter residual sequences preserved when necessary to maintain session boundaries and avoid information loss. This design allows the full recordings to be utilized while preserving meaningful temporal continuity.

The classifier begins with batch normalization of the input sequence, followed by a linear projection into a dense hidden space. An attention-based temporal aggregation module then learns adaptive importance weights over the sequence and compresses it into a context vector that highlights the most diagnostically relevant temporal information. The context vector is finally passed to a linear output layer for three-class prediction among AD, MCI, and HC. To reduce the impact of class imbalance, the final training configuration employs balanced sampling together with a focal-loss objective. Optimization is performed using AdamW, and the learning rate is adaptively updated with ReduceLROnPlateau. The final classifier configuration is selected through a hyperparameter search.

4. EXPERIMENTAND RESULTS

4.1. Experimental Setup

All experiments were conducted on a server equipped with an NVIDIA A10 GPU using Python 3.7.13 and PyTorch 1.13.1. The study was carried out in two stages. In the first stage, wav2vec2-base-960h was fine-tuned in a supervised manner to learn task-adapted representations from speech. The convolutional feature extractor of Wav2Vec2.0 was kept frozen throughout fine-tuning, while the higher-level contextual encoder, attention pooling module, auxiliary branch, and hierarchical classification heads were optimized on the target task. Optimization was performed with AdamW, using an encoder learning rate of 3×10^{-6}, dropout = 0.20, batch size = 16, gradient accumulation over 4 steps, and a maximum of 25 epochs. A cosine annealing schedule with a 0.15 warmup ratio was used, and class-specific waveform augmentation was enabled during training to improve robustness. In the second stage, the extracted 832-dimensional fusedfeatures were used to train the downstream temporal classifier. The final classifier was optimized with AdamW using a learning rate of 5×10^{-5}, batch size = 8, and dropout = 0.35. To address class imbalance, the final training configuration employed focal loss with $\gamma = 2.0$together with balanced sampling. The classifier hidden dimension was set to 384, and the attention hidden dimension was set to 420. Learning-rate adaptation was handled with ReduceLROnPlateau. The final stage-2 configuration was selected through grid search over hidden dimension, attention hidden dimension, dropout, batch size, learning rate, focal-loss setting, class-balancing strategy, and early stopping parameters.

4.2. Evaluation Protocol

Experiments were conducted under both short-speech and long-speech evaluation conditions. Evaluation was performed at the patient level, with all sessions from the same subject kept within the same split to prevent data leakage. Performance was assessed using Accuracy, macro-Precision, macro-Recall, macro-F1, and macro-ROC AUC, which together provide a balanced view of overall performance and class discrimination in the three-way classification task.

4.3. Results

Table 1 summarizes performance under the short-speech and long-speech evaluation conditions. The proposed method performed better in the long-speech setting across all reported metrics, achieving 88.04% accuracy and 87.57% macro-F1, compared with 85.19% accuracy and 85.69% macro-F1 in the short-speech setting. This result suggests that broader temporal context is beneficial for three-way cognitive-status classification.

Table 1. Performance under long-speech and short-speech evaluation conditions.

Evaluation condition	Accuracy	F1-Score (Macro)	Precision (Macro)	Recall (Macro)	ROC AUC (Macro)
Long	0.8804	0.8757	0.9083	0.8592	0.9648
Short	0.8519	0.8569	0.8717	0.8495	0.9167

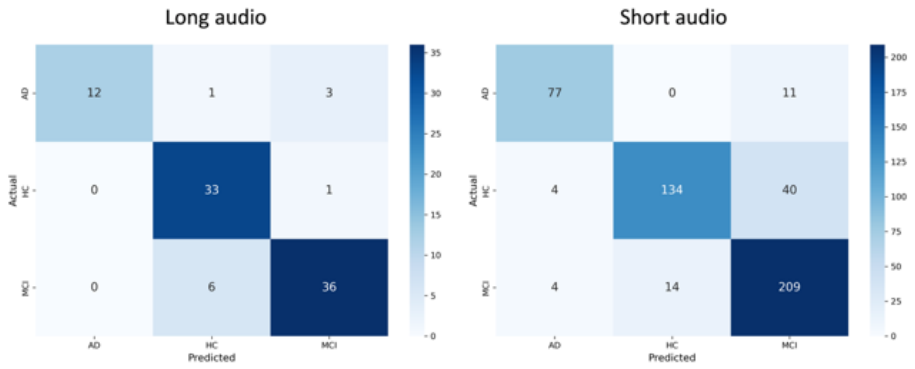

Figure 2: Confusion matrix under long-speech and short-speech evaluation conditions.

Figure 2 presents the confusion matrix for the best-performing long-speech configuration. Most residual errors occur between the neighboring clinical categories MCI and HC, whereas AD cases are recognized more reliably. This pattern is consistent with the clinical difficulty of separating subtle early impairment from healthy aging using speech alone.

Taken together, Table 1 and Figure 2 show that the long-speech condition provides a better trade-off between contextual richness and diagnostic stability than the short-speech condition. Longer recordings preserve broader temporal and prosodic structure, including cadence, pause behavior, and phrasal continuity, more effectively than short speech segments, which likely contributes to the improved performance observed under the long-speech setting.

4.4. Comparison with Prior Systems

To contextualize the proposed system, Table 2 compares our best configuration with prior speech-only studies reported on NCMMSC-AD 2021. Because earlier work does not always follow the same evaluation protocol, the table preserves the original reported metric and evaluation setting for each method rather than forcing all systems into a single directly comparable accuracy column.

Table 2. Comparison with prior speech-only methods on NCMMSC-AD 2021.

Method	Input type	Reported result(s)	Evaluation setting	Notes
Official baseline	Acoustic	Long Acc. 79.8;	Official challenge tracks	-----
[10]	Acoustic	Long Acc. 83.2;	Official challenge tracks	Raw speech; no explicit transcription stage
[12]	Acoustic	Precision 83.77; Recall 83.69; F1 83.73	AD2021/NCMMSC2021, 6 s segmented setting	Reported in F1 rather than accuracy
[13]	Acoustic	Test Acc. 85.45	Authors' own split	----
Ours	**Acoustic**	**Long Acc: 88.04**	**Patient-level hold-out**	**Wav2Vec2 + auxiliary fusion + temporal attention**

As shown in Table 2, the proposed system exceeds the official baseline and the directly comparable speech-only challenge results reported by [10]. It also compares favorably with more recent pretrained-representation approaches. However, direct numerical comparison should be interpreted with caution, since prior studies do not all use the same protocol: some report official long/short challenge-track accuracy,[12] report F1 under a 6-second segmented setting, and [13] report test accuracy on their own split.

4.5. Ablation Study

To assess the contribution of the main design choices, we conducted an ablation study in which key components of the proposed framework were removed while the remaining settings were kept fixed. Using only the auxiliary branch yielded 55.43% accuracy, showing that compact prosodic and demographic cues alone are insufficient for reliable three-way classification. Using only the acoustic branch increased accuracy to 84.78%, confirming that the supervised Wav2Vec2-based representation provides the primary discriminative signal. When acoustic and auxiliary features were fused without hierarchical supervision, accuracy further improved to 86.96%, indicating that the auxiliary cues contribute complementary information beyond the deep acoustic embedding. The full fused model achieved the best result, reaching 88.04% accuracy. These results show that the final performance depends on the joint contribution of deep acoustic representation learning, auxiliary feature fusion, and hierarchical supervision.

5. CONCLUSIONS

The proposed speech-only pipeline achieved 88.04% accuracy and 0.8757 macro-F1 on the NCMMSC-AD 2021 hold-out test set. Relative to previously reported speech-only systems on the same benchmark, this result is numerically strong and compares favorably with the official baseline. At the same time, these comparisons should be interpreted with caution because the reported evaluation settings are not fully identical across studies. Within this context, the present results suggest that task-specific fine-tuning of a pretrained acoustic encoder, together with

lightweight auxiliary fusion and temporal aggregation, can provide effective transcript-free cognitive-status classification.

Two observations are especially important. First, the stronger performance in the long-speech condition indicates that broader temporal context is beneficial for this task. Longer recordings preserve suprasegmental and discourse-level information, including cadence, pause behavior, and phrasal continuity, that may be only partially captured in shorter segments. Second, the ablation results show that the supervised Wav2Vec2-based acoustic branch provides the primary discriminative signal, while auxiliary cues contribute complementary information when fused with the acoustic representation. Using only the auxiliary branch yielded 55.43% accuracy, compared with 84.78% for the acoustic branch alone. Fusing acoustic and auxiliary features increased accuracy to 86.96%, and the full model further improved performance to 88.04%, indicating that hierarchical supervision provides an additional benefit beyond feature fusion alone.

The observed performance pattern is also consistent with the clinical structure of the task. Residual errors are concentrated more strongly between neighboring categories, especially MCI and HC, than between AD and the other groups, indicating that subtle early impairment remains more difficult to distinguish from healthy aging using speech alone. This suggests that the main challenge is no longer simply detecting advanced cognitive decline but improving sensitivity to weaker and more variable early-stage speech markers.

The present study has several limitations. First, the NCMMSC-AD 2021 dataset is modest in size and includes inter-speaker variability and session imbalance, which may affect robustness and increase the risk of dataset-specific fitting. Second, the experiments were conducted on a single benchmark, so the findings do not yet establish generalization across corpora, recording devices, or acoustic conditions. Future work should therefore examine external validation, robustness under mismatched recording settings, and uncertainty-aware subject-level prediction.

REFERENCES

[1] G. Livingston et al., "Dementia prevention, intervention, and care: 2020 report of the Lancet Commission," The lancet, vol. 396, no. 10248, pp. 413-446, 2020.

[2] N. R. Fowler et al., "Implementing early detection of cognitive impairment in primary care to improve care for older adults," Journal of internal medicine, vol. 298, no. 1, pp. 31-45, 2025.

[3] H.-L. Wang et al., "Speech silence character as a diagnostic biomarker of early cognitive decline and its functional mechanism: a multicenter cross-sectional cohort study," BMC medicine, vol. 20, no. 1, p. 380, 2022.

[4] F. Agbavor and H. Liang, "Artificial intelligence-enabled end-to-end detection and assessment of Alzheimer's disease using voice," Brain sciences, vol. 13, no. 1, p. 28, 2022.

[5] Q. Yang, X. Li, X. Ding, F. Xu, and Z. Ling, "Deep learning-based speech analysis for Alzheimer's disease detection: a literature review," Alzheimer's Research & Therapy, vol. 14, no. 1, p. 186, 2022.

[6] X. Ke, M. W. Mak, and H. M. Meng, "Automatic selection of spoken language biomarkers for dementia detection," Neural Networks, vol. 169, pp. 191-204, 2024.

[7] X. Qi, Q. Zhou, J. Dong, and W. Bao, "Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: a review," Frontiers in Aging Neuroscience, vol. 15, p. 1224723, 2023.

[8] U. Petti, S. Baker, and A. Korhonen, "A systematic literature review of automatic Alzheimer's disease detection from speech and language," Journal of the American Medical Informatics Association, vol. 27, no. 11, pp. 1784-1797, 2020.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in neural information processing systems, vol. 33, pp. 12449-12460, 2020.
[10] Y. Qin et al., "Exploiting Pre-Trained ASR Models for Alzheimer's Disease Recognition Through Spontaneous Speech," 2021.
[11] X. Zhang, W. Fu, and M. Liang, "Soft-Weighted CrossEntropy Loss for Continous Alzheimer's Disease Detection," ArXiv, vol. abs/2402.11931, 2024.
[12] L. Yang, W. Wei, S. Li, J. Li, and T. Shinozaki, "Augmented Adversarial Self-Supervised Learning for Early-Stage Alzheimer's Speech Detection," in Interspeech, 2022.
[13] X. Zhang, W. Fu, and M. Liang, "Soft-Weighted CrossEntropy Loss for Continous Alzheimer's Disease Detection," arXiv preprint arXiv:2402.11931, 2024.
[14] C. Group, "NCMMSC 2021 Alzheimer's Disease Recognition Challenge (NCMMSC-AD 2021)," presented at the National Conference on Man-Machine Speech Communication, 2021. [Online]. Available: https://github.com/THUsatlab/AD2021.

AUTHORS

Jamshaid Iqbal

Jamshaid Iqbal received his bachelor's degree in software engineering from Shenyang University of Chemical Technology, China. He is currently a master's student in Computer Science and Technology at Donghua University, Shanghai, China. His current research interests include deep learning, speech and audio processing, natural language understanding, and trustworthy artificial intelligence.

JiYun Li

Jiyun Li is currently working as Professor at the School of Information and Intelligent Science, Donghua University, and Director of the DonghuaHesheng Joint Laboratory. He is an Executive Member of the Distributed and Parallel Computing Committee of the China Computer Federation (CCF), a Member of the China Artificial Intelligence and Artificial Psychology Committee, and a Member of the China Graphics and Image Society. His current research interests include data engineering and machine learning, psychological modeling of artificial intelligence and decision-making, and the development of AI model-based platforms and applications.