Interspeech 2024
1-5 September 2024, Kos, Greece

Pre-trained Feature Fusion and Matching for Mild Cognitive Impairment Detection

Junwen Duan[1], *Fangyuan Wei*[1], *HongDong Li*[1], *Jin Liu*[1,2,*]

[1]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China; [2]Xinjiang Engineering Research Center of Big Data and Intelligent Software, School of software, Xinjiang University, Wulumuqi, China

{jwduan, weify9, hongdong, liujin06}@csu.edu.cn

Abstract

Effective diagnosis of Mild Cognitive Impairment (MCI), a preclinical stage of cognitive decline, is significant for delaying disease progression. While most current spontaneous speech-based diagnostic methods focus on English speech, the Interspeech 2024 TAUKADIAL Challenge proposed an innovative research direction to develop a language-agnostic approach to diagnose MCI. This paper proposes an MCI diagnosis method by analyzing and combining linguistic and acoustic features using the bilingual Chinese-English speech dataset provided by the challenge. We employed a pre-trained multilingual model and expressivity encoder to extract language-agnostic speech features. To overcome the challenges of data scarcity and language diversity, we implemented data augmentation and alignment to enhance the model's generalization. Our approach achieved 77.5% accuracy, demonstrating its effectiveness and potential on cross-lingual data.

Index Terms: computational paralinguistics, mild cognitive impairment, data augmentation, multimodal fusion

1. Introduction

Mild cognitive impairment (MCI) is a condition characterized by cognitive decline exceeding typical age-related changes yet not reaching the severity of dementia. This neurodegenerative condition may involve issues with memory, language, or judgment [1]. Early intervention plays a crucial role in slowing down the progression of the disease and minimizing its adverse effects on patients. Traditional diagnostic methods such as magnetic resonance imaging (MRI) and cerebrospinal fluid biomarker tests are hindered by high costs and time consumption and may only be suitable for some patients [2]. Therefore, there is an urgent need for an effective and less invasive early diagnostic method for MCI.

Research on patient speech indicates that MCI commonly leads to decreased language fluency and difficulties in word retrieval [3, 4]. Earlier studies have also shown an increase in the number and length of hesitations in patients with MCI compared to healthy controls, corresponding to a lower speech rate [5]. Consequently, some studies have shown that analyzing linguistic and paralinguistic features is an encouraging method for detecting MCI [6, 7].

Considering that previous studies have focussed on English speech data, the TAUKADIAL Challenge [8] [1] provides spontaneous speech data in both Chinese and English languages to explore the diagnosis of MCI through a cross-language generalized approach. Given the excellent performance of current multilingual pre-trained models across various downstream tasks and previous research demonstrating the effectiveness of pre-trained models in dementia detection tasks, the pre-trained features of transcriptions have become one of the features we consider [9, 10]. We utilize Prosody UnitY2 [2] to extract common prosodic features such as speech rate and pauses across different languages. Moreover, to address the scarcity of dataset samples and the diversity of languages, we aim to alleviate these issues through data augmentation and modal alignment methods, which we will elaborate on in Section 4. In this article, we present the methods used to complete the TAUKADIAL Challenge and the results submitted. By fusing and matching different modalities and language features, we achieved an accuracy of 77.5% on the test set.

2. Related work

Spontaneous speech analysis has proven effective for detecting Alzheimer's disease (AD). While language deficits in MCI resemble early AD [11, 12], MCI detection is more challenging due to its subtler decline. Significant differences exist between MCI and AD in prosody, fluency, and acoustic features [13, 3], prompting studies using acoustic, linguistic, or combined modalities for MCI diagnosis.

Acoustic speech analysis has emerged as a promising cognitive impairment detection tool. Several studies explored MCI detection using patients' prosodic and paralinguistic features [14, 13]. [15] employed correlation-based feature selection with redundant feature sets containing pause durations and personal attributes, underscoring filled pause detection's importance for effective MCI detection. Deep learning methods have shown superior performance compared to traditional methods that require substantial manual feature processing [16, 17]. The use of data augmentation to mitigate data scarcity has also been shown to be effective [18, 19]. [20] further improved the accuracy of MCI detection by proposing a novel acoustic-based deep learning approach with self-supervised clustering and online triplet generation.

Analyzing speech transcriptions is another effective MCI detection method. Studies show machine learning models can identify MCI effectively using lexical and syntactic features from tools like LIWC [21, 22]. FastText word embeddings and pre-trained model's feature representations were used for classification, showing advantages on multilingual datasets [23, 24]. ASR systems are used to automatically obtain transcriptions to overcome the scarcity of manual transcriptions [25]. These

*Corresponding author

[1]As of the submission of this paper, the referenced literature has not been published online yet. For more information, please visit https://luzs.gitlab.io/taukadial

[2]https://github.com/facebookresearch/seamless_communication

10.21437/Interspeech.2024-2386

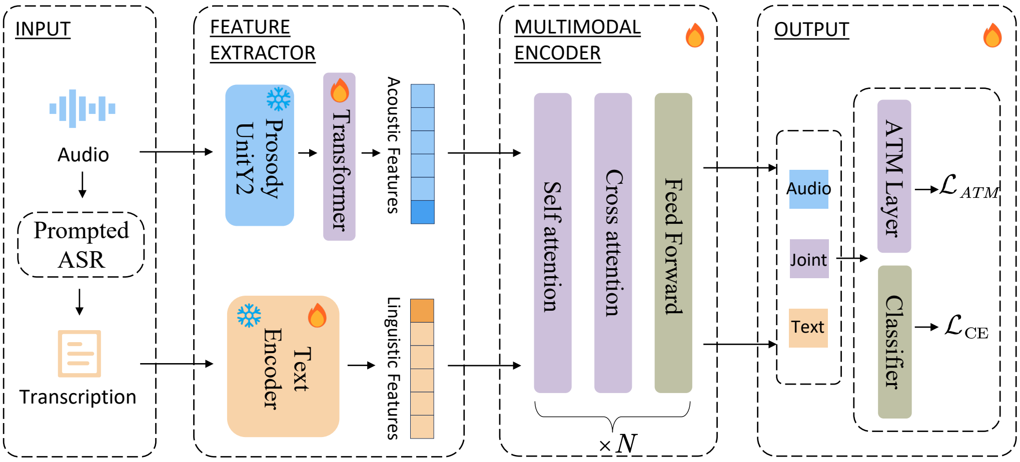

Figure 1: *The framework of the proposed method.*

studies collectively highlight the power of pre-trained models in advancing early detection of cognitive decline.

[26] and [27] utilized a combination of acoustic and linguistic features to enhance the early identification of cognitive impairments, demonstrating that both acoustic and linguistic features could effectively separate cognitive groups with high accuracy. [28] created a multifaceted speech dataset for cognitive impairment screening, integrating acoustic, lexical, and MRI features, and demonstrated a 10% F1-score improvement in MCI detection.

Certain works have achieved state-of-the-art results in Alzheimer's disease (AD) detection by utilizing advanced ASR models and larger pre-trained models [25, 29]. However, they did not attempt to fuse features from multiple modalities. Some multimodal approaches have only performed feature fusion through simplistic techniques like model voting or attention layer[30, 31]. Meanwhile, current ASR models have demonstrated excellent performance in word error rate and transcribing audio details [32], while multilingual pre-trained language models have also shown promising performance on various multilingual downstream tasks [33].

Table 1: *Statistics on number of audio segments in the dataset*

	MCI		NC	
	F	M	F	M
Train	135	87	102	63
Test	30	33	36	21
Total	165	129	138	84

3. Dataset

The dataset provided by the TAUKADIAL Challenge includes audio recordings of image descriptions from cognitively normal subjects (NC) and MCI patients. Each subject provides three audio segments describing different images in the training set. The subjects consist of native Chinese speakers and native English speakers, with the three images used in English descriptions different from those described by Chinese speakers. The test set includes audio recordings of descriptions of images in English or Chinese from different participants.

The training set comprises 186 English audio recordings (63 NC, 123 MCI) and 201 Chinese audio recordings (102 NC, 99 MCI). The test sets include 60 English audio recordings and 60 Chinese audio recordings. The dataset has been balanced in terms of age and gender.

4. Methodology

We have proposed a multimodal framework to address this challenge. As shown in Figure 1, the input audio and fine-grained transcriptions are processed by a feature extractor, with the resulting features encoded into a joint representation by a multimodal encoder. The model is then trained by optimizing different objective functions through distinct output layers.

4.1. Language-specific preprocessing

To obtain the transcriptions of audio recordings, we use the pretraining automatic speech recognition (ASR) model Whisper large-v3 [32] for transcription. This model can recognize the language present in the audio recordings and transcribe the corresponding text. In Whisper model, the *prompt* is given as a token sequence of prior information, aiming to maintain consistency in output between segments by tracking previously transcribed information.

Considering that more language pauses, filler words, and disfluencies often occur in the speech of MCI patients' speech , and current automatic speech recognition models struggle to transcribe these units accurately by default. We assign the *prompt* parameter in the official implementation of Whisper with a style of string containing pauses and filler words. For example, for English audio recording, the *prompt* could be "a pipe, um ... water pipe. hmm ... okay .. uh uh.", and for Chinese audio recording, the *prompt* could be "哦,小朋友在撈魚,誒..他們在玩什?玩...玩...。".

As a result, we can automatically obtain transcriptions that include more fine-grained information, as depicted in Figure 1 as "Prompt ASR". It should be noted that the *prompt* is used to provide a context for the language style of the transcription, and its content can be pretty diverse.

4.2. Features extraction and fusion

4.2.1. *Linguistic and acoustic feature encoding*

To better extract linguistic features from different languages, we utilize the pre-trained language model XLM-RoBERTa-base [33] to extract features from the transcription, considering that previous studies demonstrated promising performance of pre-trained models on relevant tasks [9, 29]. XLM-RoBERTa is a multilingual sentence encoder that can extract useful features for downstream tasks.

Considering that previous research has demonstrated the effectiveness of paralinguistic and acoustic features in MCI detection, directly utilizing existing ASR models to extract features may result in the loss of some segment information due to limitations in model input length. To better extract language-independent paralinguistic features such as rhythm, speech rate, and pauses, we employ Prosody UnitY2 to capture prosodic units, followed by input into a 3-layer transformer for encoding. Prosody UnitY2 is a prosody-aware speech-to-unit translation model based on the UnitY2 architecture. It aims to convey phrase-level prosody to enhance the quality and expressiveness of speech-to-speech generation.

Thus, from a same audio recording, the audio A and the text T are encoded into a sequence of embeddings $\{\boldsymbol{u_1}, ..., \boldsymbol{u_l}\}$ and $\{\boldsymbol{w_1}, ..., \boldsymbol{w_n}\}$ through the above process, with their maximum lengths set to 2048 and 514 respectively.

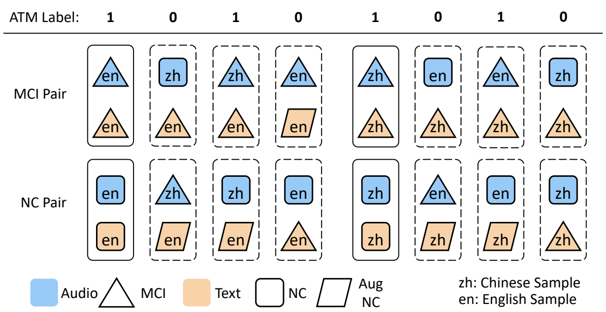

Figure 2: *Constructing pairs for ATM loss.*

4.2.2. *Multimodal fusion and data augmentation*

Our model employs various approaches to enhance the interaction between different modals. To integrate information from two modals, we utilize a 6-layer transformer as a multi-modal encoder. The output features $\{\boldsymbol{u'_1}, ..., \boldsymbol{u'_l}\}$ and $\{\boldsymbol{w'_1}, ..., \boldsymbol{w'_n}\}$ are pooled separately using mean pooling along the sequence dimension, followed by a linear transformation and a non-linear activation function:

$$\boldsymbol{t} = \tanh(\boldsymbol{W_t}\frac{1}{l}\sum_{i=1}^{l}\boldsymbol{u'_i} + \boldsymbol{b_t}) \tag{1}$$

where $\boldsymbol{W_t}$, $\boldsymbol{b_t}$ are learnable parameters. The Joint vector $\boldsymbol{J}$ is then obtained by concatenating the pooled text and image features.

Additionally, we employ audio-text matching (ATM) as a method for data augmentation and modal fusion to alleviate data scarcity. As shown in Figure 2, each audio-text pair can be a matched or mismatched pair (label=1 or label=0) based on whether the labels of the audio and text are the same. By replacing one text/audio of the pair with a text/audio sampled from the entire training set, we can augment three more audio-text pairs. Considering the significant data imbalance in the training dataset for English recordings (63 NC, 123 MCI), when sampling English NC texts, there is a 50% chance of using augmented texts generated by replacing synonyms after sampling. Specifically, we use $\boldsymbol{J}$ as the representation of the pairs to input into a linear classifier with cross-entropy loss for binary classification. Let $\boldsymbol{p}^{\mathbf{atm}}$ denote the two-class probability predicted by the classifier and $\boldsymbol{y}^{\mathbf{atm}}$ denote the one-hot ground-truth label of the pair. The ATM Loss can be defined as the cross-entropy H between $\boldsymbol{p}^{\mathbf{atm}}$ and $\boldsymbol{y}^{\mathbf{atm}}$:

$$\mathcal{L}_{\mathrm{ATM}} = \mathbb{E}_{(A,T)\sim D}\mathrm{H}(\boldsymbol{y}^{\mathbf{atm}}, \boldsymbol{p}^{\mathbf{atm}}(A,T)) \tag{2}$$

, where the order of the pairs and labels is shuffled to prevent the model from learning irrelevant sequence information.

From this, we can use four times as much augmented data as the original data to make the model learn cross-linguistic and cross-modal information through ATM.

4.3. Fine-tuning for MCI detecion

For the MCI detection task, we use cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ to fine-tune the model for classification. $\boldsymbol{J}$ are passed through a classifier consisting of two fully connected layers and a LayerNorm layer to get the final classification result. The final fine-tunning objective function is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{ATM}} \tag{3}$$

During the fine-tuning process, we froze the parameters of the Prosody UnitY2 model and the first 6 layers of the XLM-RoBERTa model. Also we use pytorch WeightedRandomSampler to load the training data to alleviate the sample imbalance. The learning rates for the cross-modal module and the classification head were set to 5 times and 10 times the encoder's learning rate, respectively.

4.4. Prediction

In the TAUKADIAL Challenge dataset, each participant provides three audio samples, and the labels of these three audio samples should be the same.

Due to the limited data, the results may vary slightly when using different seeds for parameter initialization. To improve the model's robustness and prediction accuracy, we make predictions using models trained with 5 random seeds. This approach allows us to obtain a total of 15 predictions for the same participant. The final prediction for all samples of that participant will be determined by majority voting among these results.

5. Experiments and Results

The TAUKADIAL Challenge evaluates classification tasks using specificity, sensitivity, and F1 scores and uses a balanced accuracy metric (unweighted average recall, UAR) for overall ranking. TP is the number of true positives, TN is true negatives, FP is false positives, FN is false negatives, and N is the number of samples. Then specificity $\sigma = \frac{TN}{TN+FP}$, sensitivity $\rho = \frac{TP}{TP+FN}$ and $UAR = \frac{\sigma+\rho}{2}$.

5.1. Experiment setup

As required by the challenge, we submitted five sets of classification results for evaluation on the test set. Each of our five submissions used a different method. Method 1 uses only acoustic features and a 3-layer transformer for encoding, followed

Table 2: *Test results of the classification task*

ID	Methods	Accuracy	Specificity	Sensitivity	F1	UAR
	Baseline [8]	-	-	-	-	0.592
1	Acoustic-only features	0.550	0.632	0.476	0.526	0.554
2	Text-only features	0.675	0.737	0.619	0.667	0.678
3	-w/ Prompt	0.700	0.684	0.714	0.714	0.699
4	Multimodal fusion	0.700	0.737	0.667	0.700	0.702
5	-w/ ATM	**0.775**	**0.790**	**0.762**	**0.781**	**0.776**

by a linear classifier for classification. Method 2 directly uses the ASR model to obtain transcriptions, while Method 3 inputs prompts to the model to obtain transcriptions. Both methods then use the XLM-RoBERTa-base Model with a sequence classification on top for encoding and obtaining the classification results. Methods 4 and 5 are described in Section 4, except that method 4 uses only cross-entropy loss as the objective function.

All of the above methods follow the process described in Subsection 4.3 and Subsection 4.4 for fine-tuning and prediction, using a validation set to determine hyperparameters. The validation set is constructed by holding out 10% of the training set and sampling based on subjects. The learning rate was 2e-5, the batch size was 16, and the epoch was set to 10 for all methods except method 5, which had 15 epochs.

5.2. Experiment results

The test results are shown in Table 2, where the baseline method fuses the wav2vec and eGeMAPs features. Moreover, Figure 3 illustrates the confusion matrix of the baseline method and the five submissions. The results indicate that the multimodal approach, which combines acoustic and linguistic features, outperforms unimodal methods.

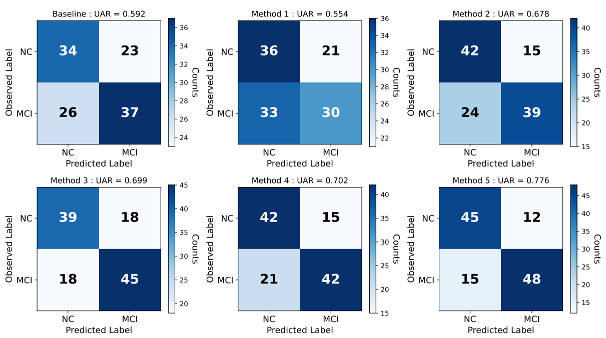

Figure 3: *Confusion matrix of test results.*

5.2.1. Unimodal methods

Acoustic-only features (Method 1) achieves a UAR of 0.554, indicating that paralinguistic cues alone are insufficient for reliable MCI diagnosis. While acoustic features can capture certain aspects of cognitive impairment, such as speech rate and pauses, they lack the semantic and contextual information provided by linguistic features.

Text-only features (Method 2) demonstrates an improved UAR of 0.678. This highlights the importance of linguistic features in detecting MCI, as language deficits are common manifestations of cognitive decline. Method 2 also achieves a higher sensitivity (0.619) and specificity (0.737) compared to Method 1, indicating its improved ability to correctly identify MCI and NC cases. However, the unimodal text-based approach still falls short of capturing the full range of cues present in speech.

Text-only features with prompts (Method 3), which incorporates prompts during ASR, achieves a UAR of 0.699, outperforming the baseline text-only method (Method 2). This improvement can be attributed to the inclusion of filler words and pauses in the transcriptions, which are known to be indicative of cognitive impairment. The prompts provided during ASR facilitate the accurate transcription of these paralinguistic features, enhancing the model's ability to capture relevant cues. This results in the model being more sensitive to MCI cases, achieving a sensitivity of 0.714. However, this also affects the specificity of the method, which is 0.684, slightly lower than Method 2.

5.2.2. Multimodal methods

Multimodal fusion (Method 4) achieves the same accuracy as Method 3, indicating that the benefits provided by multimodal fusion are limited when using only CE loss. However, Method 4 strikes a better balance between sensitivity and specificity, resulting in a slightly higher UAR (0.702).

Multimodal fusion with ATM loss (Method 5) outperforms all other methods, achieving a UAR of 0.776. This improvement can be attributed to the effective use of data augmentation and modal alignment, which mitigate the challenges of data scarcity and language diversity in the dataset. By leveraging the ATM, the model is exposed to a larger and more diverse set of audio-text pairs, enhancing its ability to learn cross-linguistic and cross-modal representations.

6. Conclusion

In this paper, we proposed an effective multimodal approach for detecting mild cognitive impairment from spontaneous speech. Our novel application of audio-text matching enabled effective data augmentation and cross-modal alignment, enhancing the model's ability to learn generalizable representations. By leveraging state-of-the-art pre-trained models and fusing acoustic and linguistic features, our method achieved promising results on the TAUKADIAL Challenge dataset, with improvements over the baseline approach. Future work will extend the method to datasets in more than two languages to further enhance its robustness and language universality.

7. Acknowledgements

This work is supported by National Science Foundation of China under Grant 62172444 and U22A2041, and Central South University Innovation-Driven Research Programme under Grant 2023CXQD018. This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

8. References

[1] B. Reisberg *et al.*, “Current evidence for subjective cognitive impairment (sci) as the pre-mild cognitive impairment (mci) stage of subsequently manifest alzheimer’s disease,” *International Psychogeriatrics*, vol. 20, no. 1, p. 1–16, 2008.

[2] C. Laske *et al.*, “Innovative diagnostic tools for early detection of alzheimer’s disease,” *Alzheimer’s & Dementia*, vol. 11, no. 5, pp. 561–578, 2015.

[3] L. Tóth *et al.*, “A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech,” *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[4] K. D. Mueller *et al.*, “Declines in connected language are associated with very early mild cognitive impairment: Results from the wisconsin registry for alzheimer’s prevention,” *Frontiers in Aging Neuroscience*, vol. 9, p. 437, 2018.

[5] G. Szatloczki *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, p. 195, 2015.

[6] K. C. Fraser *et al.*, “Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment,” *Computer Speech & Language*, vol. 53, pp. 121–139, 2019.

[7] V. Vincze *et al.*, “Linguistic parameters of spontaneous speech for identifying mild cognitive impairment and Alzheimer disease,” *Computational Linguistics*, vol. 48, no. 1, pp. 119–153, 2022.

[8] S. Luz *et al.*, “Connected speech-based cognitive assessment in chinese and english.” arXiv, 2024, DOI: 10.48550/ARXIV.2404.nnnnn (TBA).

[9] Y. Zhu *et al.*, “Wavbert: Exploiting semantic and non-semantic speech using wav2vec and BERT for dementia detection,” in *Proc. of INTERSPEECH*, 2021, pp. 3790–3794.

[10] Y. Pan *et al.*, “Using the outputs of different automatic speech recognition paradigms for acoustic- and bert-based alzheimer’s dementia detection through spontaneous speech,” in *Proc. of INTERSPEECH*, 2021, pp. 3810–3814.

[11] V. Boschi *et al.*, “Connected speech in neurodegenerative language disorders: A review,” *Frontiers in Psychology*, vol. 8, 2017.

[12] D. Beltrami *et al.*, “Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?” *Frontiers in aging neuroscience*, vol. 10, p. 369, 2018.

[13] J. J. Meilan *et al.*, “Voice markers of lexical access in mild cognitive impairment and alzheimer’s disease,” *Current Alzheimer Research*, vol. 15, no. 2, pp. 111–119, 2018.

[14] S. Kato *et al.*, “Easy screening for mild alzheimer’s disease and mild cognitive impairment from elderly speech,” *Current Alzheimer Research*, vol. 15, no. 2, pp. 104–110, 2018.

[15] G. Gosztolya *et al.*, “Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection,” in *Proc. of INTERSPEECH*, 2016, pp. 107–111.

[16] K. López-de Ipiña *et al.*, “On the analysis of speech and disfluencies for automatic detection of mild cognitive impairment,” *Neural Computing and Applications*, vol. 32, pp. 15 761–15 769, 2020.

[17] J. V. Egas-López *et al.*, “Automatic screening of mild cognitive impairment and alzheimer’s disease by means of posterior-thresholding hesitation representation,” *Computer Speech & Language*, vol. 75, p. 101377, 2022.

[18] F. Bertini *et al.*, “Automatic speech classifier for mild cognitive impairment and early dementia,” *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–11, 2021.

[19] J. Duan *et al.*, “CDA: A contrastive data augmentation method for Alzheimer’s disease detection,” in *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Jul. 2023, pp. 1819–1826.

[20] C. Qian *et al.*, “Detecting mild cognitive impairment in alzheimer’s disease using speech acoustics only: A two-stage deep metric learning approach,” in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 2682–2689.

[21] M. Asgari *et al.*, “Predicting mild cognitive impairment from spontaneous spoken utterances,” *Alzheimer’s & Dementia: Translational Research & Clinical Interventions*, vol. 3, no. 2, pp. 219–228, 2017.

[22] T. Wang *et al.*, “Towards the speech features of mild cognitive impairment: Universal evidence from structured and unstructured connected speech of chinese,” in *Proc. of INTERSPEECH*, 2019, pp. 3880–3884.

[23] H. Lindsay *et al.*, “Multilingual learning for mild cognitive impairment screening from a clinical speech task,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021, pp. 830–838.

[24] S. Amini *et al.*, “Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach,” *Alzheimer’s & Dementia*, vol. 19, no. 3, pp. 946–955, 2023.

[25] Y. Wang *et al.*, “Exploring linguistic feature and model combination for speech recognition based automatic AD detection,” in *Proc. of INTERSPEECH*, 2022, pp. 3328–3332.

[26] F. Tang *et al.*, “The joint effects of acoustic and linguistic markers for early identification of mild cognitive impairment,” *Frontiers in digital health*, vol. 3, p. 702772, 2022.

[27] G. Gosztolya *et al.*, “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.

[28] Q. Yang *et al.*, “Selecting and analyzing speech features for the screening of mild cognitive impairment,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1906–1910.

[29] J. Yuan *et al.*, “Disfluencies and fine-tuning pre-trained language models for detection of alzheimer’s disease,” in *Proc. of INTERSPEECH*, 2020, pp. 2162–2166.

[30] N. Wang *et al.*, “Modular multi-modal attention network for alzheimer’s disease detection using patient audio and language data,” in *Proc. of INTERSPEECH*, 2021, pp. 3835–3839.

[31] Y. Qiao *et al.*, “Alzheimer’s disease detection from spontaneous speech through combining linguistic complexity and (dis)fluency features with pretrained language models,” in *Proc. of INTERSPEECH*, 2021, pp. 3805–3809.

[32] A. Radford *et al.*, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[33] A. Conneau *et al.*, “Unsupervised cross-lingual representation learning at scale,” in *Proc. of ACL*, 2020, pp. 8440–8451.