Multimodal Speech Analysis for Early Detection of Mild Cognitive Impairment: A Scalable Approach

Muhammad Bilal[†*], Waleed Abdulla[*], Gary Cheung[*], Lynette Tippett[*] and Seyed Reza Shahamiri[*]
[*] University of Auckland, Auckland, New Zealand
[†]Corresponding Author: mbil956@aucklanduni.ac.nz

***Abstract*— Early detection of dementia is critical for timely intervention, yet conventional diagnostic methods remain resource intensive and limited in scalability. Speech analysis offers a non-invasive alternative, with the potential to capture early cognitive changes. In this study, we present a multimodal approach for classification of mild cognitive impairment and normal control using spontaneous speech from the TAUKADIAL Challenge dataset. Our pipeline begins with transcription via the Whisper automatic speech recognition (ASR) model and applies a novel spectrogram-level and resampling based data augmentation technique to enhance text and audio diversity while preserving semantic integrity. To represent textual information, we fine-tuned a RoBERTa-large model using Low-Rank Adaptation, enabling efficient domain adaptation on limited clinical data. Audio embeddings were extracted from the Whisper encoder, and low-level acoustic features were derived from eGeMAPS. Our best model achieved an unweighted average recall of 0.86, outperforming single-modality baselines. The results demonstrate the value of integrating ASR, Natural Language Processing, and acoustic signal processing for robust, scalable dementia screening.**

I. INTRODUCTION

Mild cognitive impairment (MCI), an at-risk state for dementia, presents subtle but measurable changes in spontaneous speech. Recent advances in machine learning, particularly in large-scale language and speech models, have significantly improved the early detection of cognitive decline. However, most existing work focuses on distinguishing Alzheimer's disease (AD) from normal control (NC) [1], leaving a gap in fine-tuned models for MCI.

Multimodal learning has shown promising results in improving diagnostic accuracy [2]. By combining heterogeneous signals such as text, audio, and acoustic features it leverages multiple modalities for better detection. In speech-based dementia screening [3], integrating textual semantics with paralinguistic and acoustic cues can provide complementary information to enhance robustness and generalization.

The TAUKADIAL Challenge [4] offers a unique opportunity to develop and evaluate such systems. While the dataset provides raw audio data, it lacks accompanying transcriptions or structured multimodal features. In this work, we develop a novel multimodal framework for early detection of dementia. Our pipeline systematically combines high- and low-level features to capture complementary information. First, we extract high-level textual representations using RoBERTa and process audio signals through the Whisper model [5]. Simultaneously, we derive low-level acoustic features using eGeMAPS [6]. These diverse features are then integrated into a unified classification framework. Our contributions are threefold:

- We propose a novel spectrogram-resampling data augmentation pipeline that enhances audio and transcript variability while preserving semantic integrity.
- We fine-tune a RoBERTa-large model using Low-Rank Adaptation **(**LoRA) on domain-specific transcripts to generate MCI sensitive text embeddings.
- We design and evaluate a language specific multimodal fusion architecture that achieves state-of-the-art performance on the English speakers in TAUKADIAL dataset.

The remainder of this paper is organized as follows: Section II reviews related work in speech-based MCI detection, Natural Language Processing (NLP) specific processing, and multimodal approaches. Section III describes the preprocessing and augmentation procedures. Section IV details our methodology, including feature extraction, model design, and fusion strategies. Section V presents experimental results and discusses performance across modalities and fusion techniques. Finally, Section VI concludes the paper and outlines future directions.

II. RELATED WORK

Recent advancements in MCI detection using spontaneous speech have gained significant attention, particularly focusing on TAUKADIAL challenge. Several approaches have been proposed, leveraging multimodal fusion, transfer learning, and universal speech representations to improve detection accuracy. However, most existing methods adopt a language-agnostic approach, potentially overlooking language-specific nuances that could enhance performance. The TAUKADIAL baseline [4]established a multilingual benchmark using English and Chinese data, achieving 59.2% unweighted average recall (UAR) with language-agnostic features. While innovative, the limited accuracy of this approach highlights the trade-offs of language independent methods.

979-8-3315-7206-8/25/$31.00

A dominant trend in recent work is the use of multilingual models to generalize across languages. Universal speech representations (wav2vec 2.0) were utilized to detect MCI using monolingual and multilingual evaluation [7]. Using monolingual evaluation, they achieved UAR of 0.75, whereas for multilingual they reported a degradation of performance to 0.72 UAR. The reported performance was moderate, which represents struggling with inter-language variability. Similarly, a multimodal approach combining acoustic and linguistic features was proposed; however, the model showed inconsistent performance across different languages [8]. The highest UAR (0.83) in the TAUKADIAL challenge was achieved using language independent classifiers [9]. Their image-specific task performance was strong; the model suffered a 32.65% performance drop (0.83 to 0.56) when using three image descriptions - falling below the 0.59 baseline. This instability demonstrates that task-specific tuning (single image) outperforms generalized approaches.

Recent work secured second position in TAUKADIAL challenge with 0.81 UAR using language-specific Whisper embeddings and ensemble modeling [10]. This approach demonstrated significant improvement over language-agnostic approaches (0.61 UAR). Their majority voting strategy showed promising results for picture description tasks in language specific settings. However, it still trails our English-optimized model. These results further validate that language-specific optimization yields superior accuracy, though the choice of feature extraction remains crucial. A multimodal fusion network indicated that language specific tuning could further improve accuracy [11]. These findings suggest that a purely multilingual approach may not fully exploit linguistic features critical for MCI detection. Recent work demonstrated that acoustic embeddings (m-VGGish) could achieve modest cross-lingual generalization using monolingual training [12]. Their audio-only approach plateaued near the baseline performance, reinforcing the need for multimodal solutions like ours that combine acoustic and linguistic cues.

Despite these advances, current approaches still face key limitations. Many models rely too heavily on multilingual [13]models that often fail to capture language-specific biomarkers. This disparity underscores the need for targeted language optimization rather than generalized approaches. Notably, studies [7], [10], [11], [12] and our work demonstrate that language-specific optimization yields superior performance, particularly for English speakers.

III. PREPROCESSING

A. Dataset

The dataset utilized in this study was obtained from TAUKADIAL challenge 2024 [4] which includes speech recordings in English and Chinese language. These recordings were collected as participants engaged in picture description tasks, administered as part of cognitive evaluations in clinical environment. For English speaker, the task involved describing three specific images: Cookie Theft, Cat Rescue and Coming & Going. Whereas Chinese speaker described three culturally themed images depicting Taiwanese culture. The training set consists of 387 recordings collected from 129 individuals, encompassing NC and MCI. The test set consists of 120 recordings from 40 individuals. Within the English language portion of the training set there are 186 audio recordings comprising of 123 MCI and 63 NC participants. This class imbalance between NC and MCI was addressed by our data augmentation pipeline shown in Fig. 1. The English test subset includes 60 recordings, equally split between MCI and NC.

B. Audio Transcription

The speech recordings of English speakers from the TAUKADIAL Challenge dataset were transcribed using the OpenAI's Whisper-Large ASR model [14]. Each recording was segmented into 30-second audio chunks with a 5-second overlap to mitigate contextual degradation and hallucination effects. To preserve edge information and remove the duplication in overlapping segments, we used semantic overlap removal step. Using the *all-MiniLM-L6-v2* model [15] from Sentence Transformers library [16], we compared the end of the previous chunk and the beginning of the current one. If the cosine similarity between overlapping word windows exceeded a threshold of 0.85, the redundant portion was trimmed from the current segment. This ensured a more concise and coherent transcription while preserving semantic continuity. Transcripts were also manually validated on 20% of the dataset to ensure fidelity, resulting in a total of 63 NC and 123 MCI labeled transcripts.

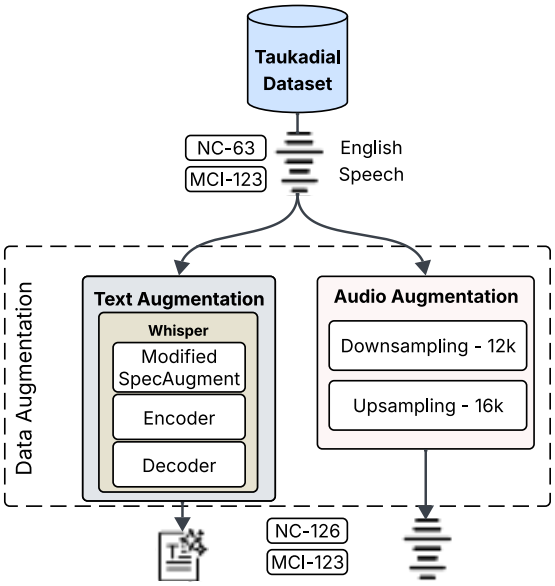

Fig. 1 Dual Modality data augmentation pipeline.
Separate augmentation techniques applied to text and audio modalities.

C. *Data Augmentation*

Although the TAUKADIAL dataset was balanced across languages, the English subset exhibits a significant class imbalance, comprising 63 NC and 123 MCI samples, which could result in low specificity. To address this, we designed a dual-modality data augmentation approach (Fig. 1), treating audio and text separately to increase variability and balance the dataset.

To address class imbalance in the NC category, a resampling-based augmentation was implemented. Original 16 kHz audio signals were downsampled to 12 kHz and then resampled back to 16 kHz [17]. This controlled transformation preserved critical speech characteristics especially as most human speech energy is concentrated below 4KHz and complies with the Nyquist sampling criterion [18]. The effectiveness of this method was verified by transcribing the augmented signals and measuring Word Error Rate (WER), which remained 4.21%, indicating high semantic similarity.

Additional experiments were performed using downsampling to 8 kHz and 10 kHz showed WERs of 8.36% and 6.25%, respectively, suggesting higher semantic distortion. Based on this, we selected 12 kHz as the optimal resampling rate and successfully doubled the NC audio samples from 63 to 126.

To further diversify the NC class in the text modality, we implemented a modified version of SpecAugment [19] prior to Whisper Encoder. Rather than relying on transcripts from resampled audio, we introduced variation at the spectrogram level to produce new text samples.

First, mel-spectrograms were processed using a global piecewise linear time warping, which randomly shifted a central time index and interpolated the spectrogram accordingly. Following warping, each spectrogram underwent time and frequency masking with dynamic widths set to 5% of the time and frequency axes, respectively. Two masks of each type were applied, and 10% of the spectrogram was protected at both temporal ends to retain contextual anchors. The warped and masked spectrograms were decoded using Whisper to generate new transcripts. Overlapping transcript regions were then semantically filtered using MiniLM-based [17], [18] similarity comparison [15], [16], ensuring unique information content across augmented samples. This process increased the NC transcript count from 63 to 126 while maintaining a WER of 6.42%.

IV. METHODOLOGY

The goal of this work is to improve NC and MCI classification performance using our novel multimodal framework that integrates text, audio, and acoustic features extracted from spontaneous speech. The overview of the proposed methodology is illustrated in Fig. 2. Due to the unavailability of pretrained language models for MCI assessment, we adapted a general-domain model using domain-specific transcripts.

A. *Transformer Fine-Tuning with LoRA*

A RoBERTa-Large model [20] was fine-tuned using LORA [21] on self-attention query/value in all 24 layers ($r = 4$, $a = 8, dropout = 0.2$; base frozen). The classification head was fine-tuned with cross-entropy loss with label smoothing. Training and evaluation were conducted using 10-fold cross-validation, stratified by speaker to prevent identity leakage. Models were trained with Cosine learning rate scheduling, 10% warmup ratio, and early stopping with patience of 20. The best fine-tuned RoBERTa model achieved a test UAR of 0.76.

B. *Audio Embedding from Whisper Encoder*

To extract deep semantic and prosodic representations from the audio modality, we leveraged the encoder block of the Whisper model [14]. Each sample was divided into 30-second chunks. For each chunk, we passed the mel-spectrogram through the Whisper encoder and extracted the final hidden state. Mean pooling was performed across all chunks per sample to produce a fixed-length embedding vector of size 1280. This embedding preserved both linguistic and acoustic attributes, offering high-level, non-handcrafted features suitable for downstream classification. Augmented audio

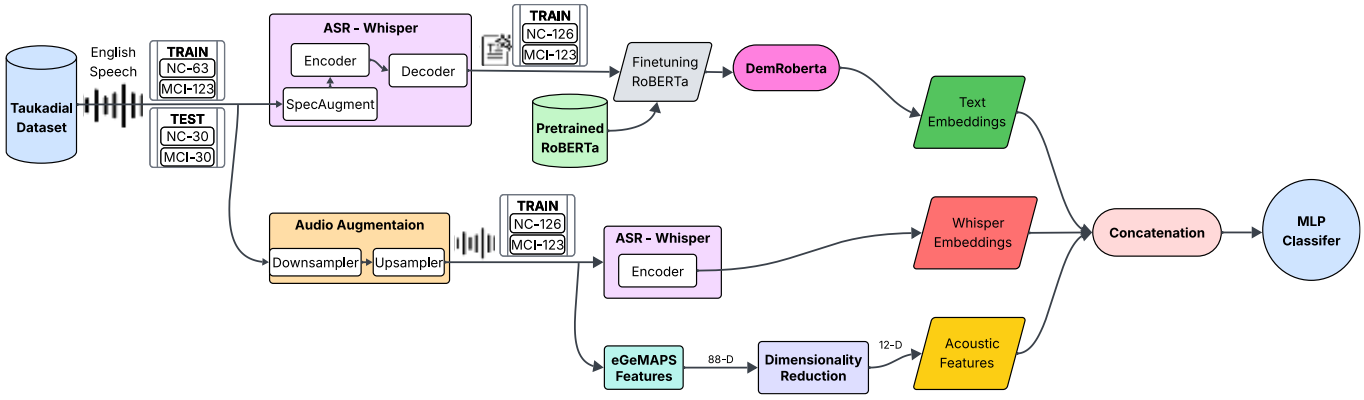

Fig. 2 Overview of proposed multimodal pipeline for MCI detection.
The framework includes preprocessing, data augmentation, acoustic feature extraction, finetuning of RoBERTa and multimodal fusion using MLP classifier.

samples from the resampling-based pipeline (Section III-B) were also processed to ensure class balance in the embedding space.

C. Acoustic Feature Extraction

To capture low-level paralinguistic features relevant to cognitive status, we extracted eGeMAPS features [6] from each audio sample. This standardized feature set includes 88 dimensions capturing temporal, spectral, and voice quality parameters [22]. Given the relatively small dataset size, high-dimensional raw features risk overfitting. To address this, we explored dimensionality reduction using Principal Component Analysis (PCA), Uniform Manifold Approximation & Projection (UMAP) [23] and PCA followed by UMAP. Performance comparisons across downstream classification tasks revealed that UMAP alone offered superior results. We tested output dimensions ranging from 2 to 88. PCA with 95% variance, reached UAR= 0.57. UMAP (k=12) performed best with UAR= 0.73. UMAP+PCA gave UAR=0.71. This preserved non-linear patterns for cognitive state discrimination.

D. Feature Extraction for Each Modality

To support multimodal learning, we extracted a unified set of vector embeddings for each audio sample, covering textual, audio, and acoustic domains. These were saved and used as inputs to the downstream classification pipeline

Textual representations were generated using a RoBERTa-large model fine-tuned with LoRA. The TextEncoder module loaded the LoRA adapter over the base transformer and computed the final embedding by averaging the CLS token outputs from the last four hidden layers. Input transcripts were tokenized using the HuggingFace tokenizer, with truncation to a maximum length of 512 tokens and no further fine-tuning during embedding generation. Using this process, we created an output vector of 1024 dimensions for each sample.

To extract audio features, we used the Whisper encoder block and introduced a chunk-level attention mechanism. Each audio file was first segmented into 30-second chunks with padding for the last chunk of each audio sample. Each chunk is passed through the Whisper encoder, and the resulting sequence embeddings were mean-pooled. The last chunk of each audio sample was selectively truncated based on valid chunk duration while ignoring the padding length. Each chunk's representation was aggregated, creating a 1280-dimensional vector that preserved both semantic and temporal information from the audio.

Low-level acoustic features were extracted using the eGeMAPS feature set [6], implemented using OpenSMILE. These features capture paralinguistic indicators such as pitch, pause duration, jitter, and shimmer [22], which are relevant for neurocognitive assessment. To reduce the dimensionality and capture non-linear structure, we applied UMAP projection on the original feature space and reduced it to 12 dimensions. The same UMAP transformation was used consistently across training and test sets.

E. Multimodal Fusion Architecture

To learn the joint representation of different modalities, we used a dynamic dimension strategy in which each modality was projected to a different dimensional space using linear layers. To make an equal contribution of embedding features, the audio embedding (1280D) was reduced to 1024D. Acoustic features (12D) projected to 128D. The concatenated features were passed through a hidden layer of 128 neurons and an MLP classifier with 2 output units.

V. Results and Discussions

To ensure balanced contribution from each modality during fusion, we applied linear projection layers to unify their dimensional scales. The audio embeddings were projected to 1024D, and the acoustic features to 128D. Multiple experiments were performed for training optimization of the MLP architecture using different numbers of hidden layers and multiple numbers of neurons in each layer using grid search. The best performance was observed with a projection of concatenated features to a 128-unit single hidden layer, striking a balance between underfitting and overfitting risk. This design choice allowed all modalities to contribute equally during multimodal fusion, leading to consistent improvements over single-modality baselines. Our best-performing configuration, using text, audio, and acoustic features, achieved the highest performance on the English subset of the TAUKADIAL dataset, to the best of our knowledge. The detailed performance metrics are summarized in Fig. 3. This represents a significant improvement over the baseline results reported in TAUKADIAL Challenge [4]. The best baseline model for English speakers used eGeMAPs features and achieved a UAR of 0.60, sensitivity of 0.80 & specificity of 0.40.

The winner of the TAUKADIAL challenge [9] achieved a UAR of 0.83 using a crosslingual model trained only on

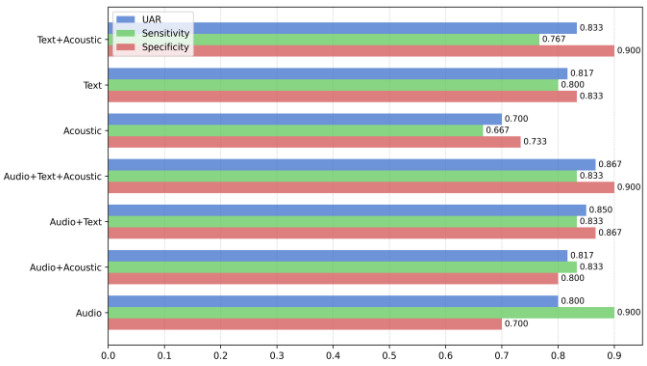

Fig. 3 Multimodal fusion performance.
Comparison of UAR, sensitivity & specificity across unimodal, bimodal & trimodal fusion strategies incorporating audio, text & acoustic modalities

image 1 (English speaker) and image 3 (Chinese speaker). These results reflect a task-specific optimization that may compromise generalization. There is a substantial performance drop when the combination of all three tasks was used (UAR: 0.56). This indicates that the study chose to exclude data diversity rather than address its impact analytically. In contrast, our approach preserved all three tasks from English speakers and achieved UAR of 0.86.

To better understand the contribution of each modality, we evaluated our model using unimodal, bimodal, and full multimodal setups. Fig. 3 presents a comparison of UAR, sensitivity and specificity across different modality combinations. The text-only model, based on LoRA-finetuned RoBERTa embeddings, achieved the highest unimodal performance with a UAR of 0.81, followed by audio-only (Whisper encoder, 0.80) and acoustic-only (UMAP-reduced eGeMAPS, 0.70). These results confirm that linguistic signals carry strong discriminative power in cognitive assessment tasks but also emphasize the additive value of acoustic and prosodic features. Bimodal combinations yielded incremental improvements. Notably, fusing text and audio achieved a UAR of 0.85, while text and acoustic reached 0.83. The full multimodal fusion of text, audio, and acoustic features achieved the highest UAR of 0.867, clearly demonstrating the complementary nature of the modalities. Additionally, the confusion matrix for the best-performing model, shown in Fig. 4, highlights balanced classification across NC and MCI classes. Sensitivity was slightly higher, suggesting that the model was well-calibrated for early cognitive decline detection, which is a desirable trait in clinical screening applications.

The proposed approach demonstrates the effectiveness of integrating pretrained models with lightweight adaptation and domain-informed augmentation for early dementia screening. The gains over TAUKADIAL baseline and previous English-only benchmarks suggest that data augmentation and balanced modality projections offer robust advantages in real-world clinical speech data. Importantly, this improvement was achieved without reliance on ensemble models, feature engineering, or language-specific tuning factors that often limit generalizability. Moreover, by aligning modalities to comparable feature spaces prior to fusion, our method simplifies the design while enhancing classification synergy.

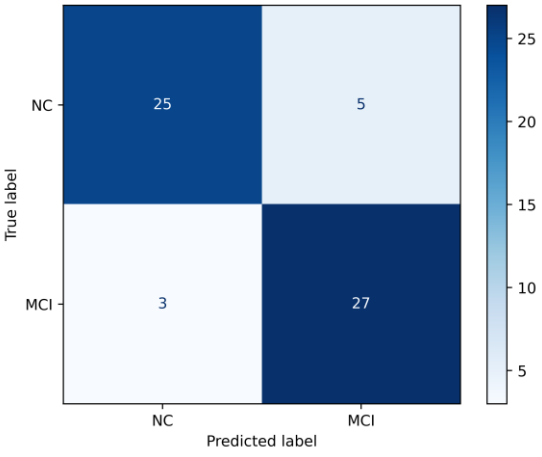

Fig. 4 Confusion matrix
Classification performance on test set of 60 samples. The model correctly classified 25 out of 30 NC samples and 27 out of 30 MCI samples, demonstrating strong discriminative ability.

VI. CONCLUSIONS

This paper presents a comprehensive multimodal framework for the detecting MCI in English speakers from spontaneous speech. Our approach, which has been proven effective in our experiments, integrates linguistic, acoustic, and semantic information using pretrained models and lightweight adaptation techniques. We proposed a spectrogram-level and resampling data augmentation pipeline to enhance transcript and audio diversity. We used LoRA for fine-tuning of RoBERTa-large language model. Moreover, we extracted meaningful audio embeddings from the Whisper encoder. Additionally, we utilized UMAP to compress clinically relevant eGeMAPS acoustic features for robust multimodal representation.

The proposed system demonstrated strong performance on the English subset of the TAUKADIAL Challenge dataset, outperforming both the official baseline and previously published English-only models. Notably, our approach achieved a UAR of 0.867 without relying on complex ensemble strategies, indicating that careful representation alignment and fusion of pretrained embeddings can be highly effective in speech-based cognitive assessment tasks.

In future work, we plan to expand our framework to support multilingual speech data, enabling generalization across languages. Finally, we intend to validate our approach on larger and more diverse clinical datasets and explore deployment pathways for real-world use in cognitive screening settings.

ACKNOWLEDGMENT

We acknowledge New Zealand eScience Infrastructure (NeSI) https://www.nesi.org.nz and DeepNet Discovery platform https://www.deepnet.auckland.ac.nz for their support

REFERENCES

[1] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer's disease detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2020, pp. 2167–2171. doi: 10.21437/Interspeech.2020-2557.

[2] P. P. Liang, A. Zadeh, and L. P. Morency, "Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and

Open Questions," *ACM Comput Surv*, vol. 56, Jun. 2024, doi: 10.1145/3656580.

[3] M. Shi, G. Cheung, and S. R. Shahamiri, "Speech and language processing with deep learning for dementia diagnosis: A systematic review," *Psychiatry Res*, vol. 329, p. 115538, Nov. 2023, doi: 10.1016/j.psychres.2023.115538.

[4] S. Luz *et al.*, "Connected Speech-Based Cognitive Assessment in Chinese and English," International Speech Communication Association, Sep. 2024, pp. 947–951. doi: 10.21437/interspeech.2024-1807.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of Machine Learning Research*, ML Research Press, 2023, pp. 28492–28518.

[6] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans Affect Comput*, vol. 7, pp. 190–202, Apr. 2016, doi: 10.1109/TAFFC.2015.2457417.

[7] A. Favaro, T. Cao, N. Dehak, and L. Moro-Velazquez, "Leveraging Universal Speech Representations for Detecting and Assessing the Severity of Mild Cognitive Impairment Across Languages," International Speech Communication Association, Sep. 2024, pp. 972–976. doi: 10.21437/interspeech.2024-2030.

[8] P. A. Pérez-Toro *et al.*, "Multilingual Speech and Language Analysis for the Assessment of Mild Cognitive Impairment: Outcomes from the Taukadial Challenge," International Speech Communication Association, Sep. 2024, pp. 982–986. doi: 10.21437/interspeech.2024-2115.

[9] B. Barrera-Altuna, D. Lee, Z. Zarnaz, J. Han, and S. Kim, "The Interspeech 2024 TAUKADIAL Challenge: Multilingual Mild Cognitive Impairment Detection with Multimodal Approach," International Speech Communication Association, Sep. 2024, pp. 967–971. doi: 10.21437/interspeech.2024-1352.

[10] F. Agbavor and H. Liang, "Multilingual Prediction of Cognitive Impairment with Large Language Models and Speech Analysis," in *Brain Sciences*, International Speech Communication Association, Dec. 2024, pp. 4211–4215. doi: 10.3390/brainsci14121292.

[11] J. Cheng, M. Elgaar, N. Vakil, and H. Amiri, "CogniVoice: Multimodal and Multilingual Fusion Networks for Mild Cognitive Impairment Assessment from Spontaneous Speech," International Speech Communication Association, Sep. 2024, pp. 4308–4312. doi: 10.21437/interspeech.2024-2370.

[12] G. Wirojburapa and D. Wanvarie, "Dementia Detection Using Transfer Learning from Recorded Speech," in *2025 17th International Conference on Knowledge and Smart Technology, KST 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 399–404. doi: 10.1109/KST65016.2025.11003329.

[13] J. Duan, F. Wei, H.-D. Li, and J. Liu, "Pre-trained Feature Fusion and Matching for Mild Cognitive Impairment Detection," International Speech Communication Association, Sep. 2024, pp. 962–966. doi: 10.21437/interspeech.2024-2386.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," https://proceedings.mlr.press/v202/radford23a.html.

[15] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, Association for Computational Linguistics, 2019, pp. 3982–3992. doi: 10.18653/v1/d19-1410.

[17] B. Vachhani, C. Bhat, and S. Kumar Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2018, pp. 471–475. doi: 10.21437/Interspeech.2018-1751.

[18] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and Trends in Signal Processing*, vol. 1, pp. 1–194, 2007, doi: 10.1561/2000000001.

[19] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2019, pp. 2613–2617. doi: 10.21437/Interspeech.2019-2680.

[20] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019.

[21] E. Hu *et al.*, "LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS," in *ICLR 2022 - 10th International Conference on Learning Representations*, International Conference on Learning Representations, ICLR, 2022.

[22] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer International Publishing, 2016. doi: 10.1007/978-3-319-27299-3.

[23] J. Healy and L. McInnes, "Uniform manifold approximation and projection," *Nature Reviews Methods Primers*, vol. 4, Dec. 2024, doi: 10.1038/s43586-024-00363-x.