MoTAS: MoE-Guided Feature Selection from TTS-Augmented Speech for Enhanced Multimodal Alzheimer's Early Screening

Yongqi Shao
Shanghai Jiao Tong University
Shanghai, China
cici_syq@sjtu.edu.cn

Bingxin Mei
Shanghai Jiao Tong University
Shanghai, China
mei18895349540@sjtu.edu.cn

Cong Tan
Shanghai Jiao Tong University
Shanghai, China
coign@sjtu.edu.cn

Hong Huo
Shanghai Jiao Tong University
Shanghai, China
huohong@sjtu.edu.cn

Tao Fang
Shanghai Jiao Tong University
Shanghai, China
tfang@sjtu.edu.cn

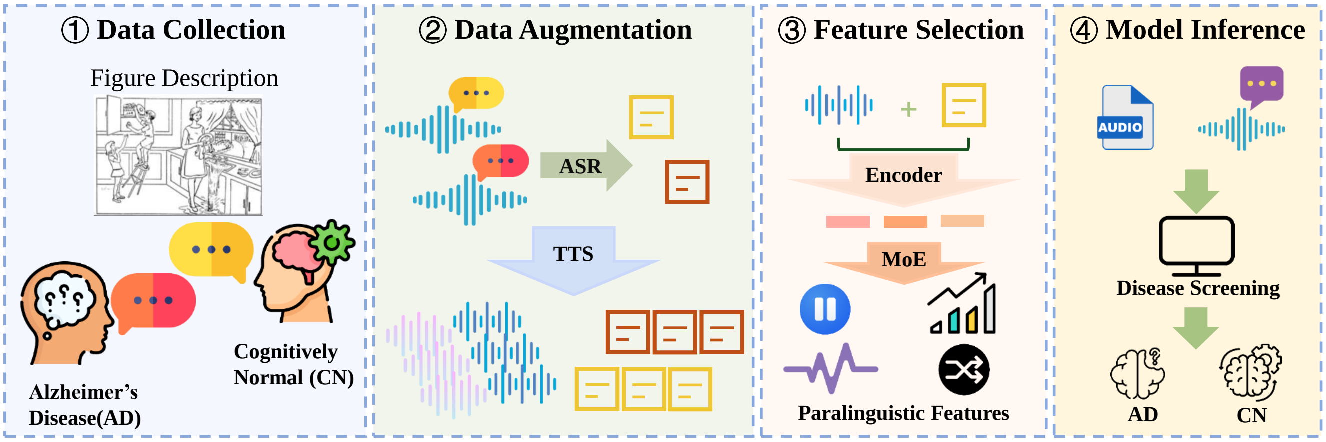

Figure 1: The Overall Pipeline of MoTAS: MoE-Guided Feature Selection from TTS-Augmented Speech for Multimodal Alzheimer's Early Screening. Some Icons Adapted from Google Material Icons (https://fonts.google.com/icons).

ABSTRACT

Early screening for Alzheimer's Disease (AD) through speech presents a promising non-invasive approach. However, challenges such as limited data and the lack of fine-grained, adaptive feature selection often hinder performance. To address these issues, we propose MoTAS, a robust framework designed to enhance AD screening efficiency. MoTAS leverages Text-to-Speech (TTS) augmentation to increase data volume and employs a Mixture of Experts (MoE) mechanism to improve multimodal feature selection, jointly enhancing model generalization. The process begins with automatic speech recognition (ASR) to obtain accurate transcriptions. TTS is then used to synthesize speech that enriches the dataset. After extracting acoustic and text embeddings, the MoE mechanism dynamically selects the most informative features, optimizing feature fusion for improved classification. Evaluated on the ADReSSo dataset, MoTAS achieves a leading accuracy of 85.71%, outperforming existing baselines. Ablation studies further validate the individual contributions of TTS augmentation and MoE in boosting classification performance. These findings highlight the practical value of MoTAS in real-world AD screening scenarios, particularly in data-limited settings.

CCS CONCEPTS

- **Applied computing → Health informatics**.

KEYWORDS

Alzheimer's Disease (AD), Speech-Based AD Screening, Text-to-Speech (TTS) Augmentation, Mixture of Experts (MoE)

ACM Reference Format:
Yongqi Shao, Bingxin Mei, Cong Tan, Hong Huo, and Tao Fang. 2025. MoTAS: MoE-Guided Feature Selection from TTS-Augmented Speech for Enhanced Multimodal Alzheimer's Early Screening. In *Proceedings of Proceedings of the 33rd ACM International Conference on Multimedia (MM '25).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746027.3755536

MM '25, October 27–31, 2025, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10…$15.00
https://doi.org/10.1145/3746027.3755536

1 INTRODUCTION

Alzheimer's disease is a progressive neurodegenerative disorder that primarily affects cognitive functions, memory, and language abilities. As the most common cause of dementia, its prevalence is rising sharply, with an estimated 55 million people affected globally. This number is projected to reach 139 million by 2050 due to aging populations[20].

The growing burden of AD poses significant challenges to healthcare systems, leading to escalating care costs, economic strain, and profound social impacts. Despite extensive research into potential treatments, no cure currently exists. This underscores the critical importance of early diagnosis in slowing disease progression and improving patient outcomes.

Traditional methods for diagnosing AD include clinical assessments, neuroimaging techniques such as magnetic resonance imaging (MRI), positron emission tomography (PET) scans, and cerebrospinal fluid (CSF) biomarker analysis. While these methods provide valuable insights into disease progression, they are expensive, require specialized resources, and are unsuitable for large-scale early screening [21, 34]. CSF testing also involves invasive lumbar puncture, causing discomfort and reducing compliance [3, 31]. Moreover, such tools often detect AD only at later stages, limiting the effectiveness of potential interventions. These limitations underscore the urgent need for a non-invasive, cost-effective, and scalable early detection approach.

Recent studies indicate that speech-based analysis is a promising alternative for AD detection, as language impairments often appear in the early stages. AD patients typically show speech features such as pauses, hesitations, reduced fluency, pronunciation errors, and lexical deficits [45]. With machine learning and deep learning models, speech analysis can automatically and objectively capture subtle linguistic and acoustic patterns linked to cognitive decline [28, 45]. Compared to conventional diagnostics, it offers a more practical solution for early screening and timely intervention.

However, despite its potential, existing speech-based AD detection methods still face several challenges[7]. The limited availability of datasets constrains the generalization ability of deep learning models across diverse populations, making them prone to overfitting. Additionally, many models treat all features equally without adaptive selection, limiting their ability to capture fine-grained cues like speech rhythm and articulation errors. Moreover, state-of-the-art deep learning approaches often require substantial computational resources, limiting their practicality for real-time clinical applications.

To address these limitations, we propose MoTAS, a speech-based Alzheimer's screening framework that leverages TTS-augmented speech and MoE-guided feature selection. Figure 1 illustrates the MoTAS pipeline. Our key contributions include:

- We propose a TTS data augmentation strategy that synthesizes speech associated with both AD and cognitively normal(CN) control groups, aiming to mitigate data scarcity and enhance model generalization.
- A MoE-guided feature selection mechanism is introduced to adaptively select features from acoustic and linguistic modalities, thereby optimizing feature utilization and reducing redundancy.
- Extensive experiments on the ADReSSo dataset demonstrate that our proposed framework significantly outperforms existing speech-based methods, achieving an accuracy of 85.71%.

By addressing challenges related to data availability, feature selection efficiency, and computational constraints, our approach advances automated, non-invasive, and scalable AD screening, providing a practical solution for early detection in real-world applications.

2 RELATED WORK

In this section, we first introduce the key advances in TTS technology for data augmentation, then discuss the role of MoE in adaptive feature selection, and finally review existing speech-based AD detection methods.

2.1 TTS for Data Augmentation in Speech Processing

In recent years, TTS technology has made significant advancements, transitioning from concatenative and parametric models to deep learning-driven approaches. Modern TTS models, such as Fish-Speech[25], VITS[22], WaveNet[40], and Tacotron[42], have greatly enhanced the naturalness and intelligibility of synthesized speech. These models utilize sequence-to-sequence architectures and advanced waveform generation techniques, enabling the production of high-quality speech with human-like prosody and articulation. As a result, TTS has been widely adopted in assistive technologies, virtual assistants, and speech synthesis research.

TTS for data augmentation has proven effective in speech-related tasks, particularly in addressing data scarcity. It has been applied in ASR[8, 44], speech emotion recognition(SER)[23, 35], and accent adaptation[8, 39], significantly improving model robustness and performance. However, its potential remains largely unexplored in AD detection. Current AD speech datasets are often small and imbalanced, making models prone to overfitting and limiting their generalizability. In this study, we leverage TTS-augmented synthetic AD speech to expand the dataset, preserving critical linguistic and acoustic features associated with cognitive decline. This enhancement boosts classification accuracy and improves the reliability of speech-based AD screening.

2.2 MoE-Guided Adaptive Feature Selection

Mixture of Experts (MoE) is a deep learning approach that improves model efficiency by dynamically selecting specialized subnetworks (experts) for different input types. It has been widely used in NLP, speech, and vision tasks to enhance scalability in large-scale learning. Recent models like Google's Switch Transformer [10] and DeepSeek-V3 [26] show that MoE can reduce computation while preserving high capacity.

In speech-related tasks, MoE has been employed to optimize feature selection and processing, leading to improved model generalization. Research has shown its effectiveness in ASR [17], speaker verification[13, 43], and SER[19, 27, 37],where it efficiently allocates different experts to process prosodic, phonetic, and spectral features, outperforming conventional deep learning approaches. Despite its success in various speech tasks, MoE has been underutilized in

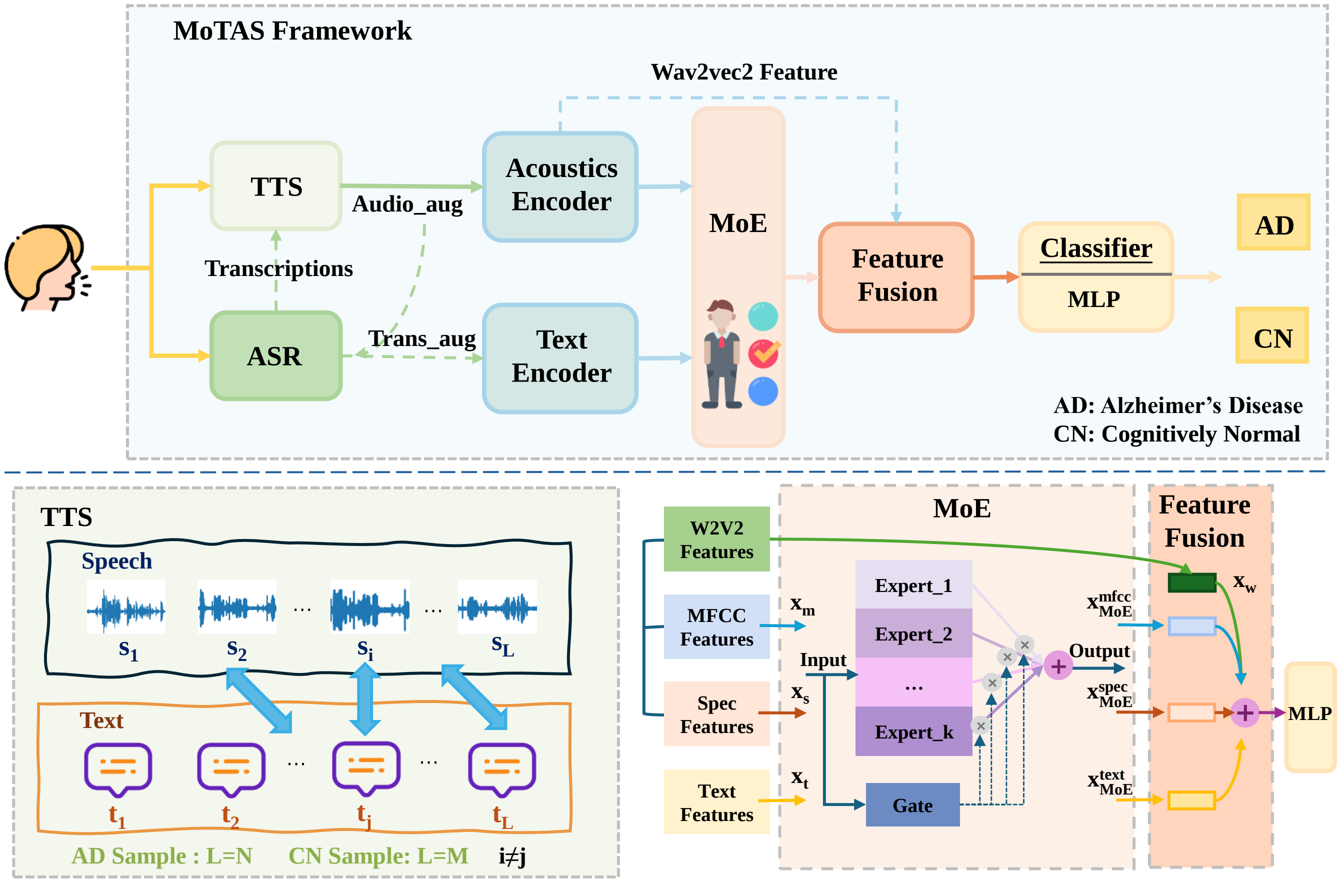

Figure 2: The Upper Part of the Diagram Illustrates the Framework of Our Proposed MoTAS; The Lower Part Details the Structure of the TTS Module on the Left and the MoE and Feature Fusion Modules on the Right.

speech-based AD detection. While existing AD classification models often process acoustic and linguistic features separately, they typically lack mechanisms to adaptively prioritize the most informative cues within each modality, such as speech rhythm, articulation errors, and lexical patterns. Furthermore, many methods treat all input features equally, which may lead to suboptimal learning and inefficient computation.

To address these limitations, we propose an MoE-guided mechanism that adaptively selects expert networks based on feature types, enabling dynamic focus on salient linguistic and paralinguistic cues. This hierarchical design enhances AD detection accuracy while reducing redundancy and computational cost.

2.3 Speech-Based AD Detection Methods

Existing speech-based AD detection methods primarily extract acoustic and linguistic features from spontaneous speech recordings and their transcriptions[5, 29, 30]. Traditional approaches rely on handcrafted features, such as speech rate, pause duration, pitch variation, and Mel-Frequency Cepstral Coefficients (MFCCs), which are then classified using Support Vector Machines (SVMs) and Random Forests[2, 4, 28]. Additionally, linguistic features derived from text transcriptions, such as lexical diversity, syntactic complexity, and word repetition patterns, have been explored to detect early cognitive decline[9, 11].

With the rise of deep learning, feature extraction has shifted from manual engineering to data-driven learning, significantly improving model performance. CNNs and RNNs have been successfully applied to Mel spectrograms and raw audio waveforms, capturing complex temporal and spectral variations[15]. Meanwhile, self-supervised learning models such as Wav2Vec2[12] enable feature extraction directly from raw speech, eliminating the need for manual feature engineering. In text analysis, Transformer-based models such as BERT[6] and DistilBERT[38] have been widely adopted to analyze transcribed speech, learning semantic coherence and syntactic changes associated with AD[32, 33]. Furthermore, multimodal models integrating speech and text features have shown superior classification performance by leveraging complementary acoustic and linguistic markers[24, 41, 46].

Despite these advancements, existing methods still face critical challenges, including limited dataset availability, suboptimal feature fusion, and high computational costs. To address these issues, we propose MoTAS, a multimodal framework that increases dataset size using TTS and enhances feature selection via a MoE mechanism. By combining synthetic speech generation with adaptive, fine-grained multimodal features selection, our approach enhances the robustness, accuracy, and scalability of automated AD detection, making it more practical for real-world deployment.

3 METHODOLOGY

The overall framework of our method is illustrated in Figure 2. Raw speech is transcribed using ASR and augmented with TTS. Both real and synthetic speech, along with their transcriptions, are encoded into multimodal features. These features are then refined through a MoE mechanism, subsequently fused, and used in the final step for classifying into AD or CN categories.

This section outlines the proposed MoTAS framework, concentrating on its key components: TTS for data augmentation, acoustic and text encoder, MoE for feature selection, and feature fusion and classification.

3.1 TTS for Data Augmentation

This study employs TTS to generate synthetic speech samples, thereby expanding the dataset while preserving disease-relevant acoustic characteristics. Since the original dataset consists solely of raw English speech recordings, we first apply ASR using Whisper [36] to obtain text transcriptions, creating paired audio-text data. The process is formulated as follows:

$$T = f_{\text{ASR}}(S), \quad \text{where } f_{\text{ASR}} = \text{Whisper} \tag{1}$$

Here, S and T represent the sets of raw speech samples and their corresponding text transcriptions, respectively. Let $S_{AD} = \{s_1^{AD}, s_2^{AD}, \ldots, s_N^{AD}\}$ and $S_{\text{CN}} = \{s_1^{\text{CN}}, s_2^{\text{CN}}, \ldots, s_M^{\text{CN}}\}$ represent the sets of AD and CN speech samples, respectively; $T_{\text{AD}} = \{t_1^{\text{AD}}, t_2^{\text{AD}}, \ldots, t_N^{\text{AD}}\}$ and $T_{\text{CN}} = \{t_1^{\text{CN}}, t_2^{\text{CN}}, \ldots, t_M^{\text{CN}}\}$ denote the corresponding ASR-transcribed texts for AD and CN speech samples, where N and M are the number of AD and CN samples in the original dataset.

Once transcriptions are obtained, a pre-trained TTS model f_{TTS} is used to synthesize new speech samples. The generated synthetic speech retains the acoustic characteristics of the reference speaker while replacing the linguistic content with transcriptions from another speaker. The synthesis process is defined as:

$$\hat{s}_i^{\text{AD}} = f_{\text{TTS}}(t_j^{\text{AD}}, s_i^{\text{AD}}), \quad \hat{s}_i^{\text{CN}} = f_{\text{TTS}}(t_j^{\text{CN}}, s_i^{\text{CN}}) \tag{2}$$

where $i = 1, \ldots, N$, $j = 1, \ldots, N$, and $i \neq j$ for AD samples, and similarly $i = 1, \ldots, M$, $j = 1, \ldots, M$, and $i \neq j$ for CN samples. Here, s_i provides the speaker identity and t_j provides the linguistic content. The TTS model synthesizes speech that combines the voice characteristics of s_i with the transcript t_j.

To augment the dataset further, each speech sample s_i can be paired with multiple transcriptions t_k from the same class (where $k \neq i$ and $k \neq j$). This intra-class pairing allows a single reference voice to be combined with various linguistic contents, producing a rich set of synthetic samples with consistent speaker identity and class label. Such augmentation improves data diversity while preserving class-specific characteristics.

In this study, we employ FishSpeech[25], a state-of-the-art TTS model designed for high-fidelity speaker-preserving synthesis, ensuring that the generated speech retains the original speaker's prosody, rhythm, and articulation. We also balance the proportion of real and synthetic samples during training to prevent overfitting.

Although the speech and text come from different speakers, the TTS model preserves key acoustic features specific to the original speaker, which is crucial for ensuring that the synthetic speech accurately reflects disease-related vocal cues. This approach allows the synthetic speech to remain authentic, accurately reflecting the acoustic traits of the original speaker, thereby enhancing the reliability and accuracy of the dataset.

The final augmented speech is defined as:

$$S_{\text{aug}} = S_{\text{orig}} \cup \{\hat{s}_i^{\text{AD}}, \hat{s}_i^{\text{CN}}\} \tag{3}$$

To prevent semantic redundancy, we rely on the fact that even when reusing textual content across speakers, the acoustic expression remains speaker-dependent due to variations in prosody and articulation. As a result, the ASR-transcribed text from real participants' speech naturally reflect speaker-specific disfluencies or omissions, introducing lexical variation that enhances diversity at both the acoustic and linguistic levels.

Thus, after generating the augmented speech samples, we perform a second round of ASR on the synthetic audio, with the resulting transcriptions serving as new textual inputs for subsequent processing. The augmented transcription process is defined as follows:

$$T_{\text{aug}} = f_{\text{ASR}}(S_{\text{aug}}), \quad \text{where } f_{\text{ASR}} = \text{Whisper} \tag{4}$$

Finally, the augmented dataset is denoted as:

$$\text{Data}_{\text{aug}} = \left\{S_{\text{aug}},\ T_{\text{aug}}\right\} \tag{5}$$

This paired multimodal dataset, together with the original dataset, serves as the input for downstream tasks including feature extraction, selection, fusion, and classification in our framework.

Overall, the TTS-augmented speech mechanism enhances dataset diversity while maintaining data quality, allowing the model to generalize more effectively across different speech patterns and mitigating overfitting issues.

3.2 Acoustic and Text Encoder

For each speech sample s_i and its corresponding text transcription t_i, we extract features from both acoustic and text modalities, forming the input feature set:

$$X = \{x_w, x_m, x_s, x_t\} = \{f_{\text{W2V2}}(s_i), f_{\text{MFCC}}(s_i), f_{\text{Spec}}(s_i), f_{\text{Text}}(t_i)\} \tag{6}$$

where $f_{\text{W2V2}}(s_i)$ represents deep phonetic features extracted using Wav2Vec2[1], capturing nuanced acoustic patterns that reflect speech prosody, phonetics, and articulation dynamics; $f_{\text{MFCC}}(s_i)$ denotes MFCC-based temporal dynamics modeled by BiLSTM[18], reflecting the spectral envelope and short-term temporal structure; $f_{\text{Spec}}(s_i)$ represents spectrogram-based features extracted using ResNet18[16], capturing the the local time-frequency energy distribution and prosodic cues; and $f_{\text{Text}}(t_i)$ corresponds to semantic and syntactic embeddings obtained from BERT[6], encoding the high-level semantic and syntactic patterns of speech.

These features comprehensively represent both the acoustic and linguistic aspects of the data, allowing MoE to selectively integrate the most relevant multimodal information.

3.3 MoE for Feature Selection

Building upon Section 3.2, we introduce the MoE mechanism to dynamically select the most informative multimodal features, optimizing classification performance. For this mechanism, we consider the following subset of features without the Wav2Vec2 component:

$$X_{\text{MoE}} = \{x_m, x_s, x_t\} = \{f_{\text{MFCC}}(s_i), f_{\text{Spec}}(s_i), f_{\text{Text}}(t_i)\} \tag{7}$$

The exclusion of $f_{\text{W2V2}}(s_i)$ is because this feature is extracted by a pre-trained model and already contains rich representational capabilities. Since it is primarily focused on representing acoustic features, further selection is not necessary. On the other hand, the text features extracted by BERT still contain valuable semantic information and are considered crucial for the task, thus they are retained in the feature selection phase. This distinction ensures that MoE can focus on integrating and enhancing the discrimination power of spectral, temporal, and semantic features.

The MoE mechanism consists of k expert networks, each designed to capture different patterns from the input feature vector. As shown in Figure 2, the three types of features are all input into the same MoE mechanism, but the MoE process for each feature is performed independently. Each expert network in the MoE mechanism produces an output for its corresponding feature type, with x_m, x_s, and x_t being input separately as follows:

$$y_i = E_i(x), \quad i \in \{1, 2, \ldots, k\} \tag{8}$$

where $E_i(\cdot)$ represents the expert network corresponding to each features.

To dynamically control the contribution of each expert, a gating network $G(\cdot)$ is employed to generate a weight vector $w \in \mathbb{R}^k$, where each element w_i corresponds to the importance of expert E_i. For each feature type $x \in \{x_m, x_s, x_t\}$, a separate gating network is used to compute the expert weights:

$$w = G(x) = \text{softmax}(W_g x + b_g) \tag{9}$$

where $W_g \in \mathbb{R}^{k \times d_x}$, $b_g \in \mathbb{R}^k$, and $d_x \in \{d_m, d_s, d_t\}$ denotes the dimension of the corresponding input feature. The softmax function ensures that the weights are positive and sum to 1, effectively determining the importance of each expert for the given input.

The final outputs of the MoE mechanism are computed separately for each feature type, providing distinct outputs for MFCC, spectrogram, and text features:

$$x_{\text{MoE}}^{mfcc} = \sum_{i=1}^{k} w_i^{\text{mfcc}} y_i^{\text{mfcc}}, \tag{10}$$

$$x_{\text{MoE}}^{spec} = \sum_{i=1}^{k} w_i^{\text{spec}} y_i^{\text{spec}}, \tag{11}$$

$$x_{\text{MoE}}^{text} = \sum_{i=1}^{k} w_i^{\text{text}} y_i^{\text{text}} \tag{12}$$

where $y_i^{\text{mfcc}}, y_i^{\text{spec}}, y_i^{\text{text}}$ are outputs from the respective expert networks for MFCC, spectrogram, and text features, and $w_i^{\text{mfcc}}, w_i^{\text{spec}}, w_i^{\text{text}}$ are the corresponding weights from the gating mechanism.

This independent processing ensures that the MoE mechanism effectively emphasizes the most relevant features for each type, enhancing the robustness and generalizability of the classification.

3.4 Feature Fusion and Classification

Following MoE-guided feature selection, we further enhance multimodal fusion by incorporating deep speech embeddings extracted via Wav2Vec2[1]. Unlike MFCC, spectrogram, and text features, which are compressed into unified representations through a modality-specific MoE mechanism, Wav2Vec2 embeddings are preserved in their raw or temporally-aggregated form. This approach leverages the pre-trained model's capacity for phonetic-level representation learning, thus avoiding unnecessary transformations and maintaining the integrity of low-level acoustic details.

To achieve comprehensive fusion of these diverse representations, we concatenate the MoE-guided features with the Wav2Vec2 feature:

$$x_{\text{final}} = \text{concat}(x_{\text{MoE}}^{mfcc}, x_{\text{MoE}}^{spec}, x_{\text{MoE}}^{text}, x_{\text{w}}) \tag{13}$$

where x_{w} represents the deep speech embeddings extracted by Wav2Vec2 in Section 3.2. The MoE-guided features provide high-level acoustic and linguistic information, while Wav2Vec2 captures phonetic and low-level speech characteristics. This dual-layered fusion strategy ensures that the final feature representation effectively integrates both high-level semantics and fine-grained acoustic details.

The multi-layer perceptron (MLP) classifier f_{MLP} consists of three fully connected layers with ReLU activations and dropout for regularization, tailored to handle complex interactions among fused features and prevent overfitting, particularly in data-limited clinical settingss:

$$f_{\text{MLP}}(x) = \text{FC}_3\left(\text{ReLU}\left(\text{Dropout}\left(\text{FC}_2\left(\text{ReLU}\left(\text{FC}_1(x)\right)\right)\right)\right)\right) \tag{14}$$

The fused multimodal representation x_{final} is subsequently passed through the MLP classifier to produce a binary classification outcome:

$$y = f_{\text{MLP}}(x_{\text{final}}) \tag{15}$$

where $y \in \{0, 1\}$ denotes the classification results (AD vs. CN). The classifier is trained using a cross-entropy loss function to optimize accuracy and robustness:

$$L = -\sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right] \tag{16}$$

Here, i represents the index of the i-th sample. By leveraging MoE for feature selection and Wav2Vec2 for deep speech embedding fusion, our approach achieves a balance between capturing discriminative multimodal information and phonetic details, thereby enhancing classification performance.

Table 1: Dataset Splits Before and After 3× Augmentation

	Train	Train_aug	Test
AD	87	253	35
CN	79	228	36
Total	166	481	71

The method design separates the roles of adaptive multimodal selection via MoE and enhancement of low-level acoustic features through Wav2Vec2. MoE targets the most discriminative features across spectral, temporal, and semantic dimensions, while Wav2Vec2 enriches the model's capability to process phonetic irregularities and subtle speech characteristics.

These enhancements are aligned with clinical observations of Alzheimer's-related speech impairments, which include semantic disorganization and phonetic irregularities such as pauses and stuttering. The fusion of MoE-guided features with Wav2Vec2 phonetic embeddings provides a robust, hierarchy-aware representation, enhancing the detection model's expressiveness. We will demonstrate the effectiveness of this integrated framework with evaluations on the ADReSSo benchmark in subsequent sections.

4 EXPERIMENTS

This section outlines the experimental setup used to assess the performance of the proposed framework. We first introduce the datasets and preprocessing steps, followed by the implementation details of model training and evaluation.

4.1 Datasets and Data Preprocessing

The dataset used in this study originates from the ADReSSo Challenge[29], which comprises English speech recordings of participants describing the "Cookie Theft" picture from the Boston Diagnostic Aphasia Exam[14]. Participants are categorized into two groups: CN and Probable AD. The original training set consists of 166 participants, while the test set includes 71 participants, with both sets balanced in terms of gender, age, and diagnostic category. Additionally, the recordings contain speech from experimenters providing instructions or engaging in brief conversations.

The original audio signals were sampled at 16kHz. During preprocessing, sentence-level timestamp annotations provided in the dataset were used to extract and segment the speech data. To ensure speaker consistency and semantic clarity, only the participant's utterances were retained. For segments shorter than 5 seconds, the original content was preserved and zero-padded to meet the target duration. Silent or invalid segments, including those with ASR failures, were excluded to maintain data quality.

All transcriptions generated by Whisper ASR were further cleaned to improve consistency and alignment across samples. This cleaning process included converting all characters to lowercase, correcting spelling errors, and filtering out non-linguistic symbols. For each cleaned speech segment, both acoustic and textual features were extracted separately. The resulting segments preserve linguistic coherence while maintaining the original acoustic structure, providing high-quality inputs for subsequent multimodal analysis.

4.2 Implementation Details

To address the limited sample size of the ADReSSo training set, we employed a speaker-consistent TTS data augmentation strategy using the Fish-Speech toolkit. Specifically, for each subject, we synthesized new speech samples by reusing transcripts from other participants while preserving the original speaker's acoustic characteristics. This approach maintains the speaker's vocal identity while introducing semantic and lexical diversity. The augmented samples were generated proportionally to the original class distribution (AD vs. CN), thereby preserving label balance. Ultimately, the training set was expanded to approximately three times its original size. To evaluate the impact of different augmentation scales on model performance, we conducted comparative experiments using training sets expended by 1.5×, 2×, and 2.5×. The optimal augmentation ratio was selected based on validation performance. A comparison of sample sizes before and after augmentation is shown in Table 1. In these ablation studies, each setting retained the original training data and supplemented the remaining portion with newly generated augmented samples as needed.

To capture acoustic characteristics at different levels, we extracted three types of acoustic features. MFCCs were computed using the Librosa library with 40 Mel filter banks, a 25 ms frame length, and a 10 ms hop size. The resulting MFCC sequences (13-dimensional per frame) were fed into a two-layer bidirectional LSTM (hidden size 128), and the final hidden state was passed through a fully connected layer to obtain fixed-length embeddings of dimension $d_m = 128$. Mel spectrograms were resized to 224×224 and processed by a pretrained ResNet18 with ImageNet weights. Segment-level features were aggregated using mean pooling, resulting in $d_s = 1000$. Phoneme-level features were obtained by averaging the last hidden states of a wav2vec2-base-960h model, yielding $d_{\mathrm{w}} = 768$.

Textual features were derived from Whisper ASR transcripts. After preprocessing, each sentence was encoded using a pretrained BERT-base model, and the [CLS] token embedding was used as the sentence-level representation, with $d_t = 1024$. All extracted features were stored for downstream alignment, fusion, and classification tasks.

For feature selection, we adopted a MoE mechanism, where each feature (MFCC, spectrogram, and text) was associated with three expert networks ($k = 3$). A shared gating mechanism dynamically assigned weights to these experts based on the input. This framework enables the model to emphasize the most discriminative features and suppress redundant information, thereby improving performance and interpretability. Notably, Wav2Vec2 features were excluded from the MoE mechanism, as they already provide high-quality phonetic representations through self-supervised pretraining and were directly incorporated in the final fusion stage.

All model components were implemented using the PyTorch framework. Training was performed using the Adam optimizer with an initial learning rate of 0.0067 and a batch size of 32. Binary cross-entropy was used as the loss function. To ensure result reliability, each experiment was repeated five times with fixed random seeds, and the final performance metrics reported represent the average across all five runs.

Table 2: Comparison of Our Method With Existing Approaches on the ADReSSo Test Set. Metrics Include Accuracy, Precision, Recall, and F1-Score, Following Definitions From the Baseline Study [29]. Our Method's Results Are Averaged Over Five Runs.

Method	Accuracy (%)	Precision (%)		Recall (%)		F1 Score (%)	
		AD	CN	AD	CN	AD	CN
Without ASR Transcripts							
ADReSSo Baseline (eGeMAPS+SVM)[29]	64.79	-	-	-	-	-	-
Wav2Vec2+TB [33]	74.65	77.42	72.50	68.57	80.56	72.73	76.32
Whisper-TL medium[24]	77.46	77.14	77.78	77.14	77.78	77.14	77.78
With ASR Transcripts							
ADReSSo Baseline (Late Fusion)[29]	78.87	77.78	80.00	80.00	77.78	78.87	78.87
WavBERT M_b (W2V2 ASR + BERT)[46]	73.24	75.00	71.79	68.57	77.78	71.64	74.67
C-Attention-Unified[41]	78.03	74.15	84.12	87.22	68.57	80.09	75.42
WavBERT M_p (W2V2 ASR + BERT + Pauses)[46]	83.10	**87.10**	80.00	77.14	**88.89**	81.82	84.21
TDNN-ASR-M5[33]	84.51	81.58	87.88	88.57	80.56	84.93	84.06
Whisper-TL-FTP Medium[24]	84.51	83.33	85.71	85.71	83.33	84.50	**84.50**
MoTAS (Ours)	**85.71**	80.49	**93.10**	**94.29**	77.14	**86.84**	84.38

5 RESULTS AND ANALYSIS

This section presents the experimental results of our MoTAS framework for speech-based Alzheimer's early screening, including comparisons with previous studies and an ablation study to evaluate the impact of key components.

5.1 Comparison with Previous Studies

We compared our proposed MoTAS framework with a range of existing speech-based AD detection models, including both acoustic-only approaches [24, 29, 33] and multimodal methods that combine speech with ASR-transcribed text [24, 29, 33, 41, 46]. The comparative results are summarized in Table 2.

As shown in the table, the proposed MoTAS framework achieves the highest overall classification accuracy (85.71%) on the ADReSSo test set, outperforming all baselines from both single- and multimodal categories. It also obtains the best CN precision (93.10%), AD recall (94.29%) and AD F1-score (86.84%), indicating strong sensitivity and balanced detection performance. These results demonstrate the effectiveness of our design in capturing AD-related speech and language patterns with greater precision and robustness.

Compared to acoustic-only models such as Wav2Vec2+TB [33] and Whisper-TL medium [24], which achieve AD recall rates of 68.57% and 77.14% respectively, our framework shows substantial improvements. For example, MoTAS increases AD recall by over 17% relative to Whisper-TL medium, while also improving accuracy and F1-score. These gains are likely attributed to the combined advantages of multimodal input, expert diversity, and MoE-guided adaptive feature selection, rather than data augmentation alone.

Among state-of-the-art multimodal systems, including WavBERT M_p [46], TDNN-ASR-M5 [33], and Whisper-TL-FTP [24], our method remains the top-performing model. Although WavBERT M_p achieves a strong AD precision of 87.10% and CN recall of 88.89%, MoTAS outperforms it across several key metrics, including AD F1-score (86.84% vs. 81.82%), AD recall (94.29% vs. 77.14%), and accuracy (85.71% vs. 83.10%). These results reflect the complementary benefits of TTS data augmentation and adaptive expert selection enabled by the MoE mechanism.

Notably, several multimodal baselines exhibit class imbalance. For example, WavBERT M_b [46] achieves only 68.57% AD recall, while C-Attention-Unified [41] shows a strong bias toward AD classification, achieving a recall of 87.22% for AD and 68.57% for CN. These outcomes suggest that naive modality fusion without adaptive control can lead to redundancy or modal dominance. In contrast, the MoE gating mechanism in our framework selectively emphasizes the most informative features for each input, improving both classification balance and model interpretability.

MoTAS also demonstrates robustness to ASR errors, which are common in spontaneous and cognitively impaired speech. The MoE gating mechanism effectively down-weights unreliable textual features, mitigating their impact on final predictions. Importantly, the high AD recall (94.29%) and F1-score (86.84%) are especially valuable in clinical screening scenarios, where reducing false negatives is critical for early diagnosis and intervention. By maintaining high sensitivity without compromising precision or overall accuracy, our model helps mitigate underdiagnosis risks.

In summary, the proposed MoTAS framework combines multimodal inputs, TTS-augmented speech, and MoE-guided adaptive feature selection to achieve balanced and superior performance across both AD and CN classes, demonstrating strong potential for real-world deployment in early-stage AD screening based on spontaneous speech.

5.2 Ablation Study

To evaluate the independent and synergistic contributions of the TTS-augmented speech data and MoE-guided feature selection mechanism in our framework, we conducted a comprehensive ablation study. As shown in Table 3, the MoE mechanism was evaluated under two conditions: without data augmentation (Experiment ID 1 and 2) and with 2× TTS augmentation, which yielded the best performance (Experiment ID 3 and 4). The TTS augmentation was

Table 3: Ablation Study on TTS Augmentation and MoE. We Evaluated Dataset Expansion at 1.5×, 2×, 2.5×, and 3×, Followed by MoE Ablation on Both the Original and Best-Performing Augmented Datasets. Results Are Averaged Over Five Runs on the ADReSSo Test Set[29].

ID	TTS (times)	MoE	Accuracy (%)	Precision (%)		Recall (%)		F1 Score (%)	
				AD	CN	AD	CN	AD	CN
1	X	X	78.28	81.46	76.06	73.71	82.86	77.20	79.20
2	X	✓	79.71	**82.58**	77.98	76.00	**83.43**	78.81	80.40
3	✓(2)	X	81.72	78.30	86.75	88.00	75.43	82.74	80.48
4	✓(2)	✓	**85.71**	80.49	**93.10**	**94.29**	77.14	**86.84**	**84.38**
5	✓(1.5)	✓	81.72	80.76	82.99	83.43	80.00	82.02	81.38
6	✓(2.5)	✓	82.86	79.29	87.73	89.14	76.57	83.87	81.68
7	✓(3)	✓	80.29	80.65	82.40	81.72	78.86	80.43	79.75

further analyzed by comparing multiple augmentation factors, including none, 1.5×, 2×, 2.5×, and 3× (Experiment ID 2, 5, 4, 6, and 7, respectively).

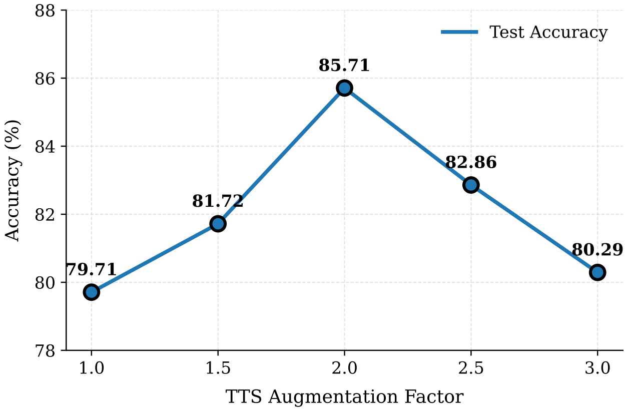

Figure 3: Test Accuracy on the ADReSSo Dataset [29] With Different TTS Augmentation Factors

The results demonstrate that the MoE mechanism consistently enhances performance across settings. Without TTS augmentation, introducing MoE increased the test accuracy from 78.28% to 79.71% (ID1 vs. ID2), indicating its effectiveness under limited data conditions. When applied to the 2× augmented dataset, MoE further improved accuracy from 81.72% to 85.71% (ID3 vs. ID4), representing a notable 3.99% gain. In addition to accuracy, other metrics such as precision, recall, and F1-score also improved significantly, confirming MoE's role in boosting robustness and discriminative capability. By dynamically weighting the importance of multimodal features, the MoE mechanism effectively reduces redundancy and enhances the model's ability to capture AD-relevant acoustic and linguistic characteristics.

For the TTS agumentation, Figure 3 illustrates the influence of varying augmentation levels on test accuracy. As the augmentation factor increased from none to 2×, the accuracy steadily improved, peaking at 85.71%. However, further augmentation to 2.5× and 3× led to a decline in accuracy to 82.86% and 80.29%, respectively. This performance degradation may be attributed to the reduced proportion of real samples, which leads the model to overfit to the synthetic distribution and impairs its generalization ability.

Therefore, effective data augmentation should not only focus on increasing quantity but also ensure the quality of synthetic data. Moderate augmentation improves sample diversity and mitigates overfitting, while excessive augmentation can negatively impact performance. Based on these findings, we selected 2× TTS augmentation as the optimal configuration, balancing dataset richness with training stability.

In summary, the ablation study validates the complementary strengths of our MoTAS framework. TTS enhances data diversity and generalization, while MoE improves the selection of discriminative features. The integration of both significantly boosts classification accuracy and robustness, providing a solid foundation for scalable and effective speech-based AD screening.

6 CONCLUSION

This study proposes an innovative framework MoTAS, which combines TTS-augmented speech data with MoE-guided feature selection to improve speech-based AD early screening. By expanding the training set with synthetic speech and adaptively selecting multimodal features, the proposed approach effectively addresses key challenges such as data scarcity, feature redundancy, and model overfitting.

Experiments on the ADReSSo dataset demonstrate that our method significantly outperforms existing speech-based models in both accuracy and robustness. The results confirm the synergistic effect of TTS augmentation and MoE-guided feature selection, which enhances model generalization while optimizing multimodal fusion under constrained computational resources.

This framework offers a flexible foundation for developing more efficient cognitive screening systems. Future work will explore its applicability to cross-lingual and cross-dataset scenarios, as well as further optimize computational efficiency to enable real-time clinical deployment. Overall, our findings highlight the potential of synthetic data generation and adaptive feature fusion in advancing early Alzheimer's screening toward more efficient, reliable, and scalable solutions.

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[2] Aparna Balagopalan and Jekaterina Novikova. 2021. Comparing acoustic-based approaches for Alzheimer's disease detection. *arXiv preprint arXiv:2106.01555* (2021).

[3] Subrato Bharati, Prajoy Podder, Dang Ngoc Hoang Thanh, and VB Surya Prasath. 2022. Dementia classification using MR imaging and clinical data with voting based machine learning models. *Multimedia Tools and Applications* 81, 18 (2022), 25971–25992.

[4] Jun Chen, Jieping Ye, Fengyi Tang, and Jiayu Zhou. 2021. Automatic detection of Alzheimer's disease using spontaneous speech only. In *Interspeech*, Vol. 2021. 3830.

[5] Xia Cui, Amila Gamage, Terry Hanley, and Tingting Mu. 2021. Identifying indicators of vulnerability from short speech segments using acoustic and textual features. *Proceedings of Interspeech 2021* (2021), 1569–1573.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[7] Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. 2024. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. *Artificial Intelligence Review* 57, 12 (2024), 325.

[8] Cong-Thanh Do, Shuhei Imai, Rama Doddipatla, and Thomas Hain. 2024. Improving Accented Speech Recognition using Data Augmentation based on Unsupervised Text-to-Speech Synthesis. In *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 136–140.

[9] Elif Eyigoz, Sachin Mathur, Mar Santamaria, Guillermo Cecchi, and Melissa Naylor. 2020. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine* 28 (2020).

[10] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[11] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2015. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's disease* 49, 2 (2015), 407–422.

[12] María Lara Gauder, Leonardo Daniel Pepino, Luciana Ferrer, and Pablo Riera. 2021. Alzheimer disease recognition using speech-based embeddings from pre-trained models. (2021).

[13] Neeraj Gaur, Brian Farris, Parisa Haghani, Isabel Leal, Pedro J Moreno, Manasa Prasad, Bhuvana Ramabhadran, and Yun Zhu. 2021. Mixture of informed experts for multilingual speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6234–6238.

[14] H Goodglass, E Kaplan, and B Barresi. 2001. Boston Diagnostic Aphasia Examination. Lippincott Williams & Wilkins. *Philadelphia, PA* (2001).

[15] Gaurav Gupta, Meghana Kshirsagar, Ming Zhong, Shahrzad Gholami, and Juan Lavista Ferres. 2021. Comparing recurrent convolutional neural networks for large scale bird species classification. *Scientific reports* 11, 1 (2021), 17085.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[17] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021), 3451–3460.

[18] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[19] Jonghwan Hyeon, Yung-Hwan Oh, Young-Jun Lee, and Ho-Jin Choi. 2024. Improving speech emotion recognition by fusing self-supervised learning and spectral features via mixture of experts. *Data & Knowledge Engineering* 150 (2024), 102262.

[20] Yun-Hee Jeon, David Foxe, Guk-Hee Suh, Huali Wang, Jacqueline C Dominguez, Rex Maukera, Sengchanh Kounnavong, and Olivier Piguet. 2024. Post-diagnosis dementia care in the Western Pacific region: assessment of needs and pathways to optimal care. *The Lancet Regional Health–Western Pacific* 50 (2024).

[21] Arun Jha and Kaushik Mukhopadhyaya. 2020. *Alzheimer's Disease: Diagnosis and Treatment Guide*. Springer Nature.

[22] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*. PMLR, 5530–5540.

[23] Siddique Latif, Abdullah Shahid, and Junaid Qadir. 2023. Generative emotional AI for speech emotion recognition: The case for synthetic emotional speech augmentation. *Applied Acoustics* 210 (2023), 109425.

[24] Jinpeng Li and Wei-Qiang Zhang. 2024. Whisper-based transfer learning for alzheimer disease classification: Leveraging speech segments with full transcripts as prompts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11211–11215.

[25] Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156* (2024).

[26] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[27] Xuan-Hao Liu, Wei-Bang Jiang, Wei-Long Zheng, and Bao-Liang Lu. 2024. MoGE: Mixture of Graph Experts for Cross-subject Emotion Recognition via Decomposing EEG. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 3515–3520.

[28] Saturnino Luz, Sofia de la Fuente, and Pierre Albert. 2018. A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919* (2018).

[29] Saturnino Luz, Fasih Haider, Sofia De la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356* (2021).

[30] Pranav Mahajan and Veeky Baths. 2021. Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. *Frontiers in Aging Neuroscience* 13 (2021), 623607.

[31] PL McGeer, H Kamo, R Harrop, EG McGeer, WRW Martin, BD Pate, and DKB Li. 1986. Comparison of PET, MRI, and CT with pathology in a proven case of Alzheimer's disease. *Neurology* 36, 12 (1986), 1569–1569.

[32] Bahman Mirheidari, Yilin Pan, Daniel Blackburn, Ronan O'Malley, and Heidi Christensen. 2021. Identifying Cognitive Impairment Using Sentence Representation Vectors.. In *Interspeech*. 2941–2945.

[33] Yilin Pan, Bahman Mirheidari, Jennifer M Harris, Jennifer C Thompson, Matthew Jones, Julie S Snowden, Daniel Blackburn, and Heidi Christensen. 2021. Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic-and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech.. In *Interspeech*. 3810–3814.

[34] Elodie Passeri, Kamil Elkhoury, Margaretha Morsink, Kerensa Broersen, Michel Linder, Ali Tamayol, Catherine Malaplate, Frances T Yen, and Elmira Arab-Tehrany. 2022. Alzheimer's disease: treatment strategies and their limitations. *International journal of molecular sciences* 23, 22 (2022), 13954.

[35] VM Praseetha and PP Joby. 2022. Speech emotion recognition using data augmentation. *International Journal of Speech Technology* 25, 4 (2022), 783–792.

[36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[37] Ali N Salman, Karen Rosero, Lucas Goncalves, and Carlos Busso. 2025. Mixture of Emotion Dependent Experts: Facial Expressions Recognition in Videos through Stacked Expert Models. *IEEE Open Journal of Signal Processing* (2025).

[38] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[39] Tian Tan, Yizhou Lu, Rao Ma, Sen Zhu, Jiaqi Guo, and Yanmin Qian. 2021. Aispeech-sjtu asr system for the accented english speech recognition challenge. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6413–6417.

[40] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* 12 (2016).

[41] Ning Wang, Yupeng Cao, Shuai Hao, Zongru Shao, and KP Subbalakshmi. 2021. Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data.. In *Interspeech*. 3835–3839.

[42] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[43] Zhiyong Wang, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Xiaopeng Wang, Yuankun Xie, Xin Qi, Shuchen Shi, Yi Lu, Yukun Liu, et al. 2025. Mixture of experts fusion for fake audio detection using frozen wav2vec 2.0. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[44] Guanrou Yang, Fan Yu, Ziyang Ma, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025. Enhancing Low-Resource ASR through Versatile TTS: Bridging the Data Gap. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[45] Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. 2022. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. *Alzheimer's Research & Therapy* 14, 1 (2022), 186.

[46] Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A Batsis, and Robert M Roth. 2021. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In *Interspeech*, Vol. 2021. 3790.