

# CoSTA: Cognitive-State-Conditioned TTS Data Augmentation Using ASR Transcripts for Alzheimer's Disease Detection

Anonymous ICME submission

**Abstract**—Speech-based Alzheimer’s Disease (AD) detection holds significant promise but is often constrained by the scarcity of pathological speech corpora. To address this bottleneck, we propose CoSTA, a Text-to-Speech (TTS)-based data augmentation framework designed to enhance AD detection. Specifically, we develop two Cognitive-State-Conditioned (CS-Cond) TTS variants by adapting an autoregressive model (CosyVoice2) and a non-autoregressive model (F5-TTS), which serve as controllable generators to synthesize speech with distinct AD-like versus Healthy-Control (HC)-like characteristics. Furthermore, to investigate whether TTS-based augmentation should be driven by Manual Transcripts (MT) or Automatic Speech Recognition (ASR)-generated ones, which may carry diagnostically relevant cues from ASR errors. Therefore, we construct a diverse transcript pool comprising MT and 36 ASR transcripts, enabling a systematic comparison between MT-driven and ASR-driven augmentation. In addition, we perform objective TTS evaluation, augmentation-factor analysis, and test-time augmentation. Extensive experiments on the ADReSS dataset demonstrate that CS-Cond TTS improves synthetic speech utility over pretrained counterparts, and that ASR-driven augmentation frequently outperforms MT-driven augmentation. Finally, CoSTA yields a 4.16% absolute improvement over the unaugmented baseline, achieving an audio-only accuracy of 85.83% on the ADReSS test set and outperforming prior methods. These findings underscore the efficacy of CoSTA for AD detection in data-scarce scenarios.

**Index Terms**—Alzheimer’s disease detection, data scarcity, text-to-speech, data augmentation, automatic speech recognition

## I. INTRODUCTION

Alzheimer’s Disease (AD), the leading cause of dementia, results in irreversible cognitive decline, including impairments in memory, language, and executive function. Due to its subtle onset, early detection is critical for timely treatment, yet conventional diagnostic methods, such as Positron Emission Tomography, are often costly and invasive [1]. Since speech production tightly couples cognition, language, and motor control, AD-related impairments often manifest in spontaneous speech as frequent pauses, temporal disfluencies, imprecise articulation, and lexical deficits [2]. Consequently, recent research has focused on developing automated speech-based AD detection models to distinguish AD from Healthy Controls (HC), positioning this approach as a non-invasive and scalable early screening tool [3]–[5]. However, robust model development is severely constrained by data scarcity, arising from ethical constraints, limited patient availability, and privacy regulations. This sparsity makes modern neural models prone to overfitting and poor generalization, highlighting the critical need for effective Data Augmentation (DA).

While prior studies have explored speech DA for AD detection [6], [7], the majority rely on traditional techniques that apply signal-level perturbations directly to the waveform, such as noise injection, pitch shifting, or time stretching. These approaches primarily generate distorted variants of existing recordings, without introducing new semantic content or explicitly modeling pathology-specific speaking traits, such as AD-like hesitations or unnatural pauses, that serve as critical discriminative cues for AD detection. Moreover, empirical evidence suggests that such indiscriminate perturbations may potentially degrade AD detection performance [7].

Recent advances in Text-to-Speech (TTS) synthesis offer a promising alternative for DA. By converting text into speech, TTS can theoretically generate an infinite number of training samples with varied acoustic characteristics, a strategy proven effective in data-scarce Automatic Speech Recognition (ASR) settings [8]. However, TTS-based DA remains largely unexplored in the context of AD detection. A critical challenge lies in the mismatch between standard TTS objectives and AD diagnostic requirements: standard TTS models are typically optimized for intelligibility and naturalness [9], [10], which inherently regularizes disfluencies and prosodic irregularities, thereby masking acoustic biomarkers of AD. Consequently, **we argue that effective DA must leverage Cognitive-State-Conditioned (CS-Cond) TTS to intentionally synthesize speech that preserves and reflects distinct AD-like versus HC-like characteristics.**

In parallel, another underexplored yet critical dimension is the choice of text sources used to drive TTS. While Manual Transcripts (MT) provides the ground truth, recent studies suggest that ASR errors may encode diagnostic cues that occasionally benefit text-level AD detection compared to perfect transcripts [11]. This raises a fundamental question for TTS-based DA: *Should synthetic speech be driven by MT or ASR transcripts?* Unlike MT, ASR transcripts introduce diverse error patterns and linguistic perturbations, which may not only increase data variance but also provide the synthetic speech with diagnostically relevant pathological cues.

To address these challenges, we propose CoSTA, a Cognitive-State-Conditioned TTS-based data Augmentation framework. Our main contributions are as follows:

- We develop two CS-Cond TTS models by adapting CosyVoice2 and F5-TTS, enabling controlled synthesis of AD-like versus HC-like speech for effective DA.
- We systematically investigate the impact of text sources for

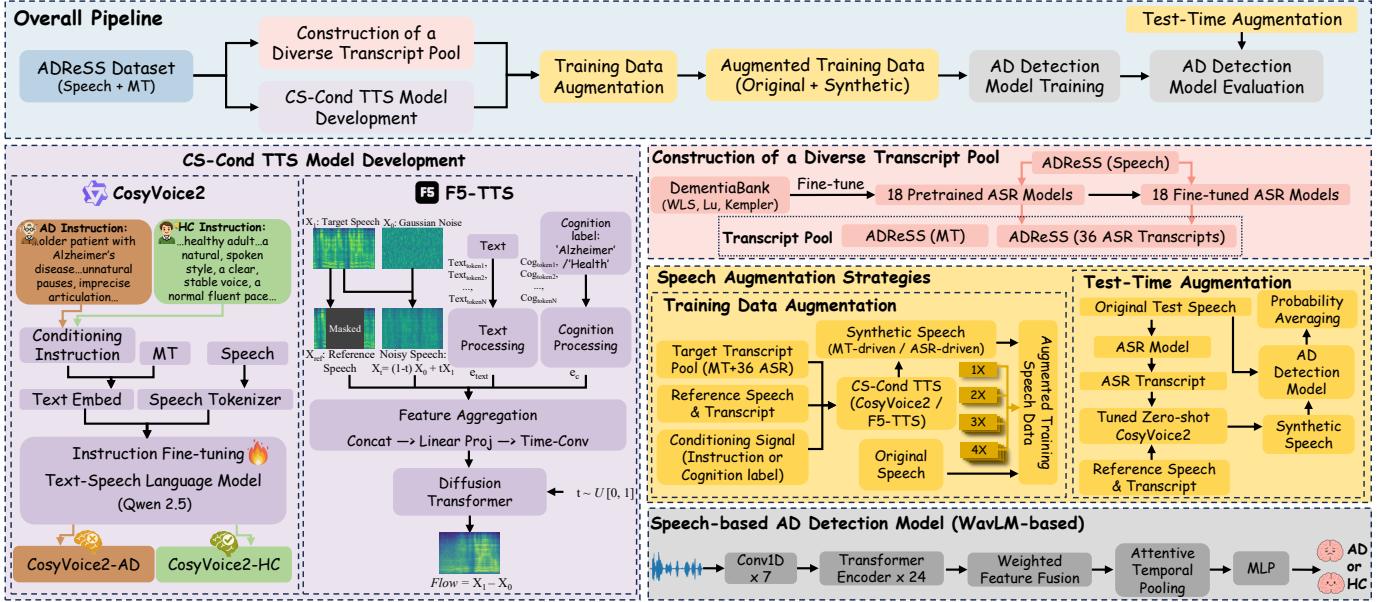


Fig. 1. Overview of the proposed CoSTA framework.

TTS by constructing a diverse transcript pool comprising MT and 36 ASR transcripts (18 each from pretrained and fine-tuned ASR models), revealing that ASR-driven augmentation frequently outperforms its MT-driven counterpart.

- Extensive experiments on the ADReSS dataset demonstrate CoSTA’s efficacy. It achieves an audio-only accuracy of 85.83% on the test set and surpassing existing baselines.

## II. METHODOLOGY

The overall framework of our method is illustrated in Fig. 1. This section outlines the four primary modules of the proposed CoSTA: CS-Cond TTS model development, construction of a diverse transcript pool, speech augmentation strategies, and the speech-based AD detection model.

### A. CS-Cond TTS Model Development

To synthesize speech with distinct AD-like or HC-like acoustic traits for DA, we adapt two advanced TTS systems.

1) *CS-Cond CosyVoice2*: CosyVoice2 [9] employs a pre-trained textual LLM (Qwen2.5-0.5B) [12] as a unified text-speech Language Model (LM) that autoregressively generates discrete speech tokens in a next-token-prediction manner. To achieve cognitive state controllability within the generation process, we leverage its instruction fine-tuning capability.

As illustrated in Fig. 1, we design two categories of natural-language instructions  $\mathcal{I}^{(c)}$  for cognitive state  $c \in \{\text{AD}, \text{HC}\}$ . During fine-tuning, an instruction is concatenated with the target MT  $y_{\text{MT}}$  using a special separator token  $<|\text{endofprompt}|>$  to form a unified text prompt:

$$y^{(c)} = [\mathcal{I}^{(c)}, <|\text{endofprompt}|>, y_{\text{MT}}] \quad (1)$$

The text embed module converts  $y^{(c)}$  into a sequence of text tokens  $\mathbf{t}^{(c)} = \{t_1, \dots, t_M\}$ . Meanwhile, the paired original speech from the training set is discretized by the speech tokenizer into  $\mathbf{s} = \{s_1, \dots, s_T\}$ .

We fine-tune the text–speech LM with teacher forcing by minimizing the negative log-likelihood of the ground-truth speech tokens  $s$  conditioned on the text prompt:

$$\mathcal{L}_{\text{LM}} = - \sum_{k=1}^T \log P_\theta(s_k \mid \mathbf{t}^{(c)}, \mathbf{s}_{<k}) \quad (2)$$

where  $\theta$  denotes the LM parameters. By performing this instruction fine-tuning on AD-specific and HC-specific subsets of the training data respectively, we obtain two specialized variants: CosyVoice2-AD and CosyVoice2-HC.

At inference time, we select a target transcript from the transcript pool and the desired cognitive instruction  $\mathcal{I}^{(c)}$  for  $c \in \{\text{AD}, \text{HC}\}$ , along with a reference speech of the corresponding cognitive category to provide the prompt speech tokens. We feed them into the corresponding model (CosyVoice2-AD or CosyVoice2-HC) to autoregressively generate target speech tokens. These tokens are subsequently processed by a flow-matching-based reconstruction module to synthesize Mel spectrograms [13], which are finally converted into the waveform via a pretrained neural vocoder HiFi-GAN [14].

2) *CS-Cond F5-TTS*: To complement the autoregressive approach, we adapt F5-TTS [10], a non-autoregressive model based on Flow Matching (FM) with a Diffusion Transformer (DiT) backbone. Unlike CosyVoice2’s text-instruction dependency, F5-TTS models speech generation as transforming a simple distribution (Gaussian noise  $x_0$ ) to the complex target speech distribution ( $x_1$ ) by estimating a velocity field  $v_t$ .

To enable cognitive state controllability, we incorporate a Cognition Processing block that is analogous to the Text Processing block, which comprises a sequence of ConvNeXtv2 [15] layers with RoPE encoding [16], to map a discrete cognition label  $l_c \in \{\text{Alzheimer}, \text{Health}\}$  into a dense embedding  $\mathbf{e}_c$ . As illustrated in Fig. 1, four types of inputs: the cognition embedding  $\mathbf{e}_c$ , the text embedding  $\mathbf{e}_{\text{text}}$ , the reference speech

mel-spectrogram  $\mathbf{x}_{\text{ref}}$ , and the noisy speech mel-spectrogram  $\mathbf{x}_t$ , are processed via a Feature Aggregation module. Specifically, these inputs are concatenated and aggregated by using a linear projection followed by convolution along the time axis. Subsequent to the feature aggregation, a DiT backbone with RoPE encoding is used to predict the flow.

The model is trained to minimize the FM loss:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,x_0,x_1} \|v_{\theta}(t, x_t, \mathbf{C}) - (x_1 - x_0)\|^2 \quad (3)$$

where  $\theta$  denotes the model parameters,  $t \sim \mathcal{U}[0, 1]$  is the time step,  $x_t = (1-t)x_0 + tx_1$  represents the interpolated noisy state, and  $\mathbf{C}$  represents the aggregated conditioning context formed by the text, reference speech, and cognition features. Unlike the training strategy of CosyVoice2, we train the unified CS-Cond F5-TTS model on a mixture of AD and HC samples.

During the inference stage, we select a target transcript from the transcript pool and specify a cognition label  $l_c \in \{\text{Alzheimer, Health}\}$ . We then pair these with a reference speech and its corresponding reference transcript of the same cognitive category selected from the training set. These inputs are fed into CS-Cond F5-TTS to generate log mel spectrograms, which are subsequently converted into audio signals via the pretrained neural vocoder Vocos [17].

### B. Construction of a Diverse Transcript Pool via ASR Models

To systematically investigate the impact of input text sources on TTS-based DA, we construct a diverse transcript pool comprising both MT and 36 ASR transcripts. Specifically, we employ the Transformers library to fine-tune 18 pre-trained ASR models across four widely used families: Wav2Vec2 [18], HuBERT [19], WavLM [20], and Whisper [21]. These models were selected for their diverse architectures and training paradigms to yield a wide range of transcription fidelities and error distributions. The selected models, available on HuggingFace, include: ***wav2vec2***-{base-100h, base-960h, large-960h, large-960h-lv60, large-960h-lv60-self, large-xlsr-53-english, xls-r-1b-english}, ***hubert***-{large-ls960-ft, xlarge-ls960-ft}, ***wavlm***-libri-clean-100h-{base-plus, large}, and ***whisper***-{tiny, base, small, medium, large, large-v2, large-v3}. After fine-tuning, we utilize the full ensemble of 36 ASR models (18 pre-trained and 18 fine-tuned) to transcribe the AD dataset, generating 36 distinct ASR transcripts for each speech sample. Subsequently, these ASR transcripts, alongside the ground-truth MT, are fed into the TTS models as text inputs, which augments the

TABLE I  
OBJECTIVE EVALUATION RESULTS ON THE TTS TEST SET.  
↓: LOWER THE BETTER.

TTS model	Variant	MCD ↓	log-F0 ↓	FAD ↓
CosyVoice2	Pretrained-AD	6.854	0.328	8.542
	CS-Cond-AD	<b>5.436</b>	<b>0.305</b>	<b>2.192</b>
	Pretrained-HC	7.537	0.328	6.206
	CS-Cond-HC	<b>6.933</b>	<b>0.301</b>	<b>2.459</b>
F5-TTS	Pretrained-AD	6.187	0.320	2.964
	CS-Cond-AD	<b>5.734</b>	<b>0.310</b>	<b>2.545</b>
	Pretrained-HC	7.340	0.316	3.165
	CS-Cond-HC	<b>7.271</b>	<b>0.312</b>	<b>2.744</b>

original training dataset and enables a comparative analysis between ASR-driven and MT-driven augmentation strategies.

### C. Speech Augmentation Strategies

1) *Training Data Augmentation*: We leverage the four TTS models (pretrained and CS-Cond variants of CosyVoice2 and F5-TTS) and the transcript pool (MT and 36 ASR transcripts) to augment the training speech data. Formally, let the original AD detection dataset be denoted as  $\mathcal{D} = \mathcal{D}_{\text{AD}} \cup \mathcal{D}_{\text{HC}}$ . The AD subset is defined as  $\mathcal{D}_{\text{AD}} = \{(s_i^{\text{AD}}, t_i^{\text{AD}})\}_{i=1}^M$ , where  $s$  represents the original speech and  $t$  denotes the MT for  $M$  subjects. Similarly, the HC subset is  $\mathcal{D}_{\text{HC}} = \{(s_j^{\text{HC}}, t_j^{\text{HC}})\}_{j=1}^N$  for  $N$  subjects. For each speech sample  $s_k$  (where  $k$  indexes any sample in  $\mathcal{D}$ ), we have access to a set of 37 transcripts  $\mathcal{T}_k = \{t_k\} \cup \{t_k^{(a)}\}_{a=1}^{36}$ , comprising the single MT and 36 ASR transcripts. To generate a synthetic sample for a specific target transcript  $t_{\text{tar}} \in \mathcal{T}_k$  using a TTS model  $\Phi$ , we select the conditioning signal  $C_k$  corresponding to the sample's ground-truth label and the synthesis process is formulated as:

$$\hat{s}_k = \Phi(t_{\text{tar}}, C_k, s_{\text{ref}}, t_{\text{ref}}) \quad (4)$$

where  $s_{\text{ref}}$  and  $t_{\text{ref}}$  denote the reference speech and text, respectively. To achieve variable augmentation factors (e.g., 2×, 3×, 4×), we use two strategies:

- Self-Reference Synthesis (2×): We set  $s_{\text{ref}} = s_k$  and  $t_{\text{ref}} = t_k$ . This generates a synthetic variant  $\hat{s}_k$  that retains the original speaker's timbre while incorporating the linguistic characteristics of the target transcript  $t_{\text{tar}}$ .
- Intra-Class Cross-Synthesis (> 2×): To further diversify the training distribution, we randomly sample a reference speech  $s_n$  from a different subject  $n$  within the same label (i.e.,  $\text{label}(k) = \text{label}(n)$ ,  $k \neq n$ ). This produces synthetic samples combining the linguistic content of subject  $k$  with the timbre of subject  $n$ . This strategy allows us to construct datasets beyond 2× by iteratively adding cross-synthesized samples.

2) *Test-Time Augmentation (TTA)*: In addition to training-time augmentation, we implement TTA to enhance performance. Since the ground-truth cognitive labels are unavailable during testing, the CS-Cond generation cannot be explicitly applied. Instead, we fine-tune a zero-shot CosyVoice2 model on the training set without cognitive instruction conditioning. For a given test speech  $s_{\text{test}}$ , we first transcribe it using the same ASR model employed during training augmentation. We then synthesize a variant  $\hat{s}_{\text{test}}$  using the tuned zero-shot model with  $s_{\text{test}}$  as the reference. Both the original  $s_{\text{test}}$  and the synthetic  $\hat{s}_{\text{test}}$  are fed into the AD detection model to obtain probability distributions  $\mathbf{P}_{\text{ori}} = [p_{\text{HC}}^{\text{ori}}, p_{\text{AD}}^{\text{ori}}]$  and  $\mathbf{P}_{\text{syn}} = [p_{\text{HC}}^{\text{syn}}, p_{\text{AD}}^{\text{syn}}]$ , respectively. The final prediction is derived via probability averaging:  $\mathbf{P}_{\text{final}} = (\mathbf{P}_{\text{ori}} + \mathbf{P}_{\text{syn}})/2$ .

### D. Speech-based AD Detection Model

We propose an audio-only AD detection model built upon a pretrained WavLM. Given a raw speech waveform sampled at 16 kHz, a 7-layer strided convolutional feature extractor first converts the sample-level signal into a sequence of frame-level representations with an overall temporal downsampling factor

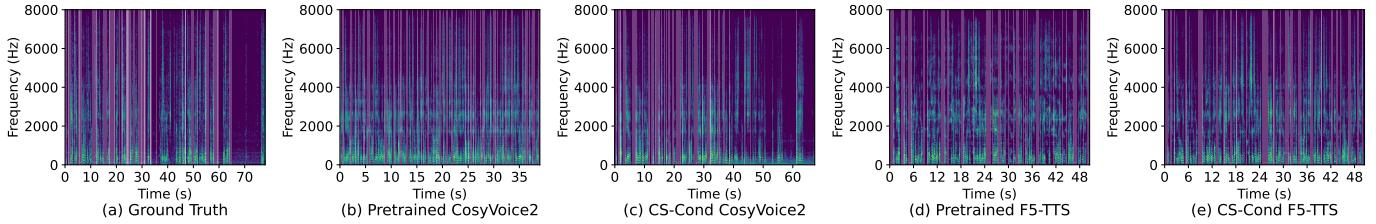


Fig. 2. Spectrograms of the ground-truth speech and synthetic speech generated by the four TTS models for an AD manual transcript from the TTS test set.

of 320. The resulting features are normalized and linearly projected to a 1024-dimensional embedding space, which is then processed by a 24-layer Transformer encoder. To fully exploit hierarchical representations across Transformer layers, we apply a weighted feature fusion module that computes a learnable, softmax-normalized combination of hidden states from all Transformer layers at each time step. The fused frame-level features are subsequently aggregated into a fixed-dimensional representation using attentive temporal pooling. Finally, this pooled representation is fed into a lightweight three-layer MLP classifier followed by a softmax function to predict the probability of the binary labels (AD vs. HC). The entire model is trained end-to-end using the cross-entropy loss.

### III. EXPERIMENTS

### A. Dataset

For AD detection, we used the ADReSS dataset [22], which contains a training set of 108 subjects (54 AD, 54 HC;  $\approx 1.7$  hours) and a test set of 48 subjects (24 AD, 24 HC;  $\approx 0.9$  hours). Each subject performed the “Cookie Theft” picture description task, producing paired speech-manual transcript pairs. We randomly partitioned the ADReSS training set into a TTS training set (45 AD, 45 HC) and a TTS test set (9 AD, 9 HC) at a 5:1 ratio. Based on the provided timestamp information, we segmented the original samples into speech-text pairs of approximately 30 seconds for TTS model training. To fine-tune the ASR models, we utilized three DementiaBank subsets (WLS, Lu, Kempler) comprising 245 samples ( $\approx 3$  hours) of the identical picture description task.

### B. Implementation Details

For training CosyVoice2, we used the Adam optimizer with a lr of  $1 \times 10^{-5}$  and adopted a dynamic batch strategy, with a maximum of 2,000 frames per batch. We empirically found that training for one epoch was sufficient. For training F5-TTS, we used the AdamW optimizer with a lr of  $1 \times 10^{-5}$  and a batch size of 8, training for 40 epochs. We employed three metrics to objectively evaluate the TTS models: Mel Cepstral Distortion (MCD) [23], Log  $F_0$  Root Mean Square

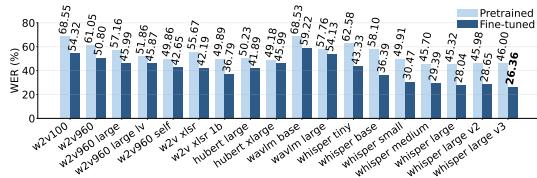


Fig. 3. Mean WER (%) of 36 ASR models on ADReSS dataset.

Error ( $\text{Log-}F_0$  RMSE) [24], and Frechet Audio Distance (FAD) [25]. Specifically, we synthesized speech using the MT from the TTS test set. These metrics were calculated by pairing the synthesized speech with the ground-truth speech. For the calculation of FAD, we utilized WavLM embeddings. For fine-tuning ASR models, we used the AdamW optimizer with a lr of  $1 \times 10^{-5}$  and a batch size of 8, training for 20 epochs, utilizing Word Error Rate (WER) for evaluation. For training the AD detection model, we utilized AdamW optimizer with a lr of  $5 \times 10^{-5}$  and a batch size of 8, training for 30 epochs, reporting the average accuracy obtained from five independent runs for evaluation. All experiments were conducted on NVIDIA A800 GPUs with 80GB of VRAM. We will release the code upon acceptance.

#### IV. RESULTS AND ANALYSIS

#### A. Objective Evaluation of TTS Models

Table I presents the objective evaluation results of the TTS models. From this table, we observe that: **1)** Across all metrics for both AD and HC classes, the CS-Cond models consistently outperform their pretrained counterparts. This confirms that explicitly modeling the cognitive state brings the acoustic profile significantly closer to the ground truth. **2)** CS-Cond CosyVoice2 outperforms CS-Cond F5-TTS across all metrics for both categories, suggesting superior distribution matching.

Fig. 2 visualizes the spectrograms of synthesized speech for an AD subject. We can observe that the pretrained models tend to generate fluent speech, whereas the CS-Cond models exhibit prosodic characteristics much closer to those of AD patients, such as frequent pauses (spectral gaps). These visualizations corroborate the quantitative results in Table I.

### *B. Diversity of ASR Transcripts*

Fig. 3 illustrates the WER of the 18 ASR models, comparing their pretrained and fine-tuned versions on the ADReSS dataset. As observed, fine-tuning consistently improves ASR

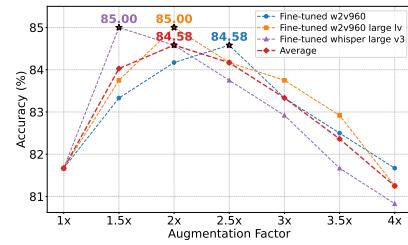


Fig. 4. Impact of augmentation factor on AD detection accuracy. Experiments utilize CS-Cond CosyVoice2 to synthesize speech based on the three high-performing ASR transcripts (identified in **blue** in Table II).

TABLE II

AD DETECTION ACCURACY (%) COMPARISON. THE BASELINE IS TRAINED ON THE ORIGINAL TRAINING SET (108 SAMPLES), WHILE COMPARISON MODELS USE A 2× AUGMENTED TRAINING SET (ORIGINAL + SYNTHETIC, 216 SAMPLES). SYNTHETIC SPEECH IS GENERATED BY FOUR TTS MODELS USING EITHER MANUAL TRANSCRIPTS (MT) OR 36 ASR TRANSCRIPTS (18 PRETRAINED + 18 FINE-TUNED). ↑ INDICATES ACCURACY SURPASSING THE BASELINE (81.67%). **BOLD** DENOTES ASR-DRIVEN AUGMENTATION OUTPERFORMING THE MT-DRIVEN COUNTERPART PER TTS MODEL. **BLUE BOLD** VALUES DENOTE CONFIGURATIONS (>84%) SELECTED FOR FURTHER ANALYSIS ON AUGMENTATION FACTORS AND TEST-TIME AUGMENTATION.

Baseline (Original training set)		81.67							
		— 2× TTS Data Augmentation —							
TTS model Text source		Pretrained CosyVoice2		CS-Cond CosyVoice2		Pretrained F5-TTS		CS-Cond F5-TTS	
Manual transcripts (MT)		80.83		82.50 <sub>↑0.83</sub>		81.25		82.08 <sub>↑0.41</sub>	
ASR transcripts		Pretrained	Fine-tuned	Pretrained	Fine-tuned	Pretrained	Fine-tuned	Pretrained	Fine-tuned
w2v100		<b>81.67</b>	<b>81.67</b>	82.08 <sub>↑0.41</sub>	82.50 <sub>↑0.83</sub>	80.42	80.83	81.25	81.67
w2v960		<b>81.67</b>	80.83	81.67	<b>84.17</b> <sub>↑2.50</sub>	79.58	<b>81.67</b>	81.67	<b>82.92</b> <sub>↑1.25</sub>
w2v960 large		<b>82.50</b> <sub>↑0.83</sub>	<b>82.08</b> <sub>↑0.41</sub>	82.08 <sub>↑0.41</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>81.67</b>	<b>81.67</b>	81.67	<b>83.33</b> <sub>↑1.66</sub>
w2v960 large lv		<b>81.25</b>	80.00	82.08 <sub>↑0.41</sub>	<b>85.00</b> <sub>↑3.33</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>83.33</b> <sub>↑1.66</sub>	<b>83.75</b> <sub>↑2.08</sub>
w2v960 self		<b>81.25</b>	<b>81.25</b>	81.67	<b>82.92</b> <sub>↑1.25</sub>	80.83	<b>82.50</b> <sub>↑0.83</sub>	81.25	<b>82.92</b> <sub>↑1.25</sub>
w2v xlsr		80.00	80.42	81.25	<b>83.33</b> <sub>↑1.66</sub>	80.42	81.25	80.42	<b>82.50</b> <sub>↑0.83</sub>
w2v xlsr 1b		80.83	<b>81.25</b>		82.08 <sub>↑0.41</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>81.67</b>	<b>82.50</b> <sub>↑0.83</sub>	81.67
hubert large		<b>81.67</b>	80.00	<b>83.33</b> <sub>↑1.66</sub>	<b>82.92</b> <sub>↑1.25</sub>	81.25	<b>81.67</b>	81.25	<b>82.50</b> <sub>↑0.83</sub>
hubert xlarge		<b>82.50</b> <sub>↑0.83</sub>	80.42	<b>83.75</b> <sub>↑2.08</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>81.67</b>	<b>82.50</b> <sub>↑0.83</sub>	<b>82.50</b> <sub>↑0.83</sub>	<b>83.33</b> <sub>↑1.66</sub>
wavlm base		80.42	<b>82.08</b> <sub>↑0.41</sub>	81.25	<b>82.92</b> <sub>↑1.25</sub>	80.83	<b>82.08</b> <sub>↑0.41</sub>	82.08 <sub>↑0.41</sub>	<b>83.33</b> <sub>↑1.66</sub>
wavlm large		<b>81.25</b>	<b>82.08</b> <sub>↑0.41</sub>	<b>83.33</b> <sub>↑1.66</sub>	<b>83.33</b> <sub>↑1.66</sub>	<b>82.08</b> <sub>↑0.41</sub>	<b>82.08</b> <sub>↑0.41</sub>	<b>83.33</b> <sub>↑1.66</sub>	<b>82.50</b> <sub>↑0.83</sub>
whisper tiny		78.33	80.42	81.67	81.67	80.83	<b>82.08</b> <sub>↑0.41</sub>	<b>82.92</b> <sub>↑1.25</sub>	81.67
whisper base		80.00	<b>81.67</b>	80.00	<b>82.92</b> <sub>↑1.25</sub>	81.25	<b>82.50</b> <sub>↑0.83</sub>	81.67	<b>82.92</b> <sub>↑1.25</sub>
whisper small		80.42	80.00	81.67	80.42	<b>82.08</b> <sub>↑0.41</sub>	<b>82.08</b> <sub>↑0.41</sub>	81.67	<b>82.50</b> <sub>↑0.83</sub>
whisper medium		80.42	80.83	82.08 <sub>↑0.41</sub>	<b>83.33</b> <sub>↑1.66</sub>	<b>82.08</b> <sub>↑0.41</sub>	81.25	<b>82.50</b> <sub>↑0.83</sub>	81.67
whisper large		80.42	<b>83.33</b> <sub>↑1.66</sub>	82.08 <sub>↑0.41</sub>	<b>83.33</b> <sub>↑1.66</sub>	80.83	<b>82.08</b> <sub>↑0.41</sub>	80.83	<b>82.92</b> <sub>↑1.25</sub>
whisper large v2		<b>81.25</b>	<b>82.08</b> <sub>↑0.41</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>82.92</b> <sub>↑1.25</sub>	<b>82.50</b> <sub>↑0.83</sub>	<b>83.33</b> <sub>↑1.66</sub>
whisper large v3		80.83	<b>81.25</b>	<b>82.92</b> <sub>↑1.25</sub>	<b>84.58</b> <sub>↑2.91</sub>	81.25	<b>81.67</b>	<b>82.50</b> <sub>↑0.83</sub>	<b>83.37</b> <sub>↑1.70</sub>
Ratio (2×Augmentation > Baseline)		7/37		28/37		16/37		24/37	
Ratio (ASR-driven > MT-driven)		19/36		20/36		23/36		22/36	

performance, yielding lower WER across all models. The WER values span a broad range, from a maximum of 68.55% to a minimum of 26.36%. This substantial variance in transcription quality provides a diverse pool of ASR transcripts for the subsequent TTS-based DA.

### C. AD Detection Accuracy with CS-Cond TTS Augmentation

We compared the AD detection accuracy of the baseline model (trained only on the original data) against models trained with 2× augmentation using different TTS models and text sources. The results are summarized in Table II. From these results, we observe that: **1)** CS-Cond models demonstrate superior augmentation efficacy compared to their pretrained counterparts. Specifically, regarding the ratio of text sources (MT + 36 ASR transcripts) that yield accuracy surpassing the baseline, CS-Cond CosyVoice2 achieves a success rate of 28/37, significantly outperforming the 7/37 of pretrained CosyVoice2. Similarly, CS-Cond F5-TTS achieves 24/37 compared to 16/37 for pretrained F5-TTS. This demonstrates that the cognitive state-aware speech generated by our CS-Cond models is more effective for DA. **2)** ASR-driven augmentation frequently outperforms its MT-driven counterpart. For each TTS model, more than half of the 36 ASR-driven configurations yield higher accuracy than using MT (specifically 19/36, 20/36, 23/36, and 22/36). We attribute this to the nature of the text sources: MT provides perfect transcriptions, resulting in synthetic speech that highly overlaps linguistically with the original speech. In contrast, ASR transcripts contain

recognition errors that are often non-random. They likely reflect pathological acoustic features of AD patients, such as articulation slurring. These errors are preserved during speech synthesis, thereby increasing the diversity of the training data and enhancing AD detection performance.

### D. Impact of Augmentation Factor on AD Detection Accuracy

We investigated the impact of various augmentation factors (ranging from 1× to 4×) on AD detection accuracy by employing the Intra-Class Cross-Synthesis strategy. The results are illustrated in Fig. 4. We observe that the performance follows an inverted-U curve, with the optimal range lying between 1.5× and 2.5×. On average, the best performance is achieved at an augmentation factor of 2× (corresponding to a 1:1 mixture of original and synthetic data). Excessive augmentation factors lead to performance degradation. This is likely because, while the synthetic speech is realistic, it inevitably contains artifacts. An excessively high proportion of synthetic data causes the detection model to overfit to the generative features of the TTS system rather than the genuine pathological characteristics.

TABLE III  
EFFECTIVENESS OF TTA ON AD DETECTION ACCURACY (%).

TTS text source	w/o TTA	w/ TTA
Fine-tuned w2v960	84.17	85.42 <sub>↑1.25</sub>
Fine-tuned w2v960 large lv	85.00	<b>85.83</b> <sub>↑0.83</sub>
Fine-tuned whisper large v3	84.58	85.42 <sub>↑0.84</sub>
Average	84.58	85.56 <sub>↑0.98</sub>

TABLE IV  
COMPARISON WITH TRADITIONAL DA AND PREVIOUS STUDIES.

Method (Traditional DA)	Accuracy (%)	Method (Previous Studies)	Accuracy (%)
Baseline (WavLM-based)	81.67	Whisper + MLP [26]	79.17
+ Noise Addition	82.50	Wav2Vec2 + Linear [27]	80.83
+ Pitch Shifting	79.17	AW-HuBERT [28]	81.67
+ Time Stretching	82.08	CoSTA (Ours)	<b>85.83<sup>†4,16</sup></b>

### E. Effectiveness of Test-Time Augmentation (TTA)

Finally, we applied TTA to the three ASR text sources from Fig. 4 under the  $2 \times$  augmentation setting. The results are presented in Table III. We observe that after implementing TTA, the accuracy for all configurations improved by approximately 1%, ultimately achieving a best accuracy of 85.83%.

### F. Comparison with Traditional DA and Previous Studies

Table IV presents a comprehensive comparison. On the left, we observe that traditional DA methods (e.g., noise addition, time stretching) yield only marginal improvements (0.4%–0.8%) or even degradation (pitch shifting) over the baseline. On the right, compared to previous audio-only studies [26]–[28], our proposed CoSTA achieves a significantly higher accuracy of 85.83%, surpassing the baseline by 4.16% and demonstrating the effectiveness of CS-Cond TTS for DA.

## V. CONCLUSIONS

In this paper, we proposed CoSTA, a novel DA framework that leverages CS-Cond TTS to address the inherent data scarcity challenge in speech-based AD detection. Our extensive experiments yield three key insights. First, explicitly conditioning TTS models on cognitive states enables the synthesis of speech with realistic pathological prosody (e.g., pause patterns), significantly enhancing downstream augmentation utility compared to standard pretrained TTS. Second, we demonstrate that ASR-driven augmentation frequently yields superior performance compared to MT-driven augmentation. This suggests that non-random ASR errors likely encode diagnostically relevant linguistic perturbations that are effectively preserved during synthesis, thereby enriching the diversity and robustness of the training data. Finally, by integrating TTA, CoSTA achieves an audio-only accuracy of 85.83% on the ADReSS test set, surpassing the unaugmented baseline by 4.16% and outperforming prior methods, thereby validating the effectiveness of CoSTA. Future work will explore cross-lingual pathological speech synthesis and leverage synthesized data to train unified speech understanding models for AD speech.

## REFERENCES

- [1] Ellen Elisa De Roeck, Peter Paul De Deyn, and Eva Dierckx, “Brief cognitive screening instruments for early detection of Alzheimer’s Disease: a systematic review,” *Alzheimer’s research & therapy*, 2019.
- [2] Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling, “Deep learning-based speech analysis for alzheimer’s disease detection: a literature review,” *Alzheimer’s Research & Therapy*, 2022.
- [3] Yuanchao Li, Zixing Zhang, Jing Han, Peter Bell, and Catherine Lai, “Semi-supervised cognitive state classification from speech with multi-view pseudo-labeling,” in *ICASSP*, 2025.
- [4] Jinchao Li, Yuejiao Wang, Junan Li, Jiawen Kang, Bo Zheng, et al., “Detecting neurocognitive disorders through analyses of topic evolution and cross-modal consistency in visual-stimulated narratives,” *IEEE Journal of Selected Topics in Signal Processing*, 2025.
- [5] Yin-Long Liu, Yuanchao Li, Rui Feng, Liu He, Jia-Xin Chen, et al., “Leveraging cascaded binary classification and multimodal fusion for dementia detection through spontaneous speech,” in *Interspeech*, 2025.
- [6] Lior Hatson and Sri Krishnan, “Spontaneous speech feature analysis for alzheimer’s disease screening using a random forest classifier,” *Frontiers in Digital Health*, 2022.
- [7] Dominika Woszczyk, Anna Hlédiková, Alican Akman, and Soteris Demetriou, “Data augmentation for dementia detection in spoken language,” in *Interspeech*, 2022.
- [8] Wing-Zin Leung, Heidi Christensen, and Stefan Goetze, “Text-to-dysarthric-speech generation for dysarthric automatic speech recognition: is purely synthetic data enough?,” in *SPECOM*, 2025.
- [9] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, et al., “Cosyvoice 2: Scalable streaming speech synthesis with large language models,” *arXiv preprint arXiv:2412.10117*, 2024.
- [10] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, et al., “F5-TTS: A fairytales that fakes fluent and faithful speech with flow matching,” in *ACL*, 2025.
- [11] Changye Li, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov, “Useful blunders: Can automated speech recognition errors improve downstream dementia classification?,” *Journal of Biomedical Informatics*, 2024.
- [12] Qwen Team, “Qwen2. 5: A party of foundation models, september 2024,” URL <https://qwenlm.github.io/blog/qwen2>, 2024.
- [13] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, and Maximilian Nickel, “Flow matching for generative modeling,” in *ICLR*, 2023.
- [14] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *NeurIPS*, 2020.
- [15] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, and Zhuang Liu, “Convnext v2: Co-designing and scaling convnets with masked autoencoders,” in *CVPR*, 2023.
- [16] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu, “Roformer: Enhanced transformer with rotary position embedding,” *Neurocomputing*, 2024.
- [17] Hubert Siuzdak, “Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis,” in *ICLR*, 2024.
- [18] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “Wav2Vec 2.0: A framework for self-supervised learning of speech representations,” *NeurIPS*, 2020.
- [19] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, and Ruslan Salakhutdinov, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE Transactions on Audio, Speech, and Language Processing*, 2021.
- [20] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, 2022.
- [21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavy, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *ICML*, 2023.
- [22] Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney, “Alzheimer’s dementia recognition through spontaneous speech: The ADReSS challenge,” in *Interspeech*, 2020.
- [23] John Kominek, Tanja Schultz, and Alan W Black, “Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion.,” in *SLTU*, 2008.
- [24] Cheng-Cheng Wang, Zhen-Hua Ling, Bu-Fan Zhang, and Li-Rong Dai, “Multi-layer f0 modeling for hmm-based speech synthesis,” in *ISCSLP*, 2008.
- [25] Kevin Kildour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms,” in *Interspeech*, 2019.
- [26] Yifan Gao, Long Guo, and Hong Liu, “Leveraging multimodal methods and spontaneous speech for Alzheimer’s Disease identification,” in *ICASSP*, 2025.
- [27] Yin-Long Liu, Rui Feng, Jia-Hong Yuan, and Zhen-Hua Ling, “Clever hans effect found in automatic detection of alzheimer’s disease through speech,” in *Interspeech*, 2024.
- [28] Zhiqiang Guo and Zhenhua Ling, “Exploring the topics of audio words for detecting alzheimer’s disease from spontaneous speech,” *IEEE Signal Processing Letters*, 2023.