Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Beyond Manual Transcripts: The Potential of Automated Speech Recognition Errors in Improving Alzheimer's Disease Detection

Yin-Long Liu[1], *Rui Feng*[1], *Jia-Xin Chen*[1], *Yi-Ming Wang*[1], *Jia-Hong Yuan*[1,2*], *Zhen-Hua Ling*[1,2]

[1]National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, P. R. China
[2]Interdisciplinary Research Center for Linguistic Sciences, University of Science and Technology of China, Hefei, P. R. China

lyl2001@mail.ustc.edu.cn, {jiahongyuan, zhling}@ustc.edu.cn

Abstract

Recent breakthroughs in Automatic Speech Recognition (ASR) have enabled fully automated Alzheimer's Disease (AD) detection using ASR transcripts. Nonetheless, the impact of ASR errors on AD detection remains poorly understood. This paper fills the gap. We conduct a comprehensive study on AD detection using transcripts from various ASR models and their synthesized speech on the ADReSS dataset. Experimental results reveal that certain ASR transcripts (ASR-synthesized speech) outperform manual transcripts (manual-synthesized speech) in detection accuracy, suggesting that ASR errors may provide valuable cues for improving AD detection. Additionally, we propose a cross-attention-based interpretability model that not only identifies these cues but also achieves superior or comparable performance to the baseline. Furthermore, we utilize this model to unveil AD-related patterns within pre-trained embeddings. Our study offers novel insights into the potential of ASR models for AD detection.

Index Terms: AD detection, ASR errors, cross attention

1. Introduction

Alzheimer's Disease (AD), the leading cause of dementia, is a progressive neurodegenerative disorder causing irreversible brain damage and cognitive decline in memory, language, attention, and executive function [1]. Early detection is crucial, and compared to traditional clinical methods, speech-and-language-based automatic AD diagnosis has gained increasing attention for its non-invasive, cost-effective, and convenient nature [2–7], further advanced by recent challenges [8–11].

Studies show that linguistic features often outperform acoustic features alone in AD detection [12, 13]. However, manual transcripts is labor-intensive, making it impractical for large-scale datasets. In contrast, Automatic Speech Recognition (ASR) offers efficient, rapid transcripts, reducing manual effort. Recent ASR advancements, such as Wav2Vec2 [14], HuBERT [15], WavLM [16], and Whisper [17], have achieved remarkable performance, facilitating fully automated AD detection using ASR transcripts. Although previous studies suggest ASR transcripts may underperform compared to manual ones [18, 19], a comprehensive evaluation of various ASR models for AD detection is limited [20, 21]. This study addresses that gap by fine-tuning 18 variants of the four ASR models on three datasets from DementiaBank (DB) and evaluating them on the ADReSS challenge dataset [8], generating 36 ASR transcripts with varying Word Error Rate (WER) for each sample (18 from the original models and 18 from the fine-tuned ones).

Pre-trained language models like BERT [22] can capture complex patterns for detecting cognitive decline and have proven effective in AD detection [23, 24]. We extract BERT embeddings from ASR and manual transcripts of the ADReSS dataset and design a self-attention-based model for AD detection. Our results show that certain ASR transcripts outperform manual ones, suggesting that ASR errors may offer valuable cues for distinguishing AD from Healthy Controls (HC). To further explore this, we synthesize speech from ASR and manual transcripts using the cutting-edge Text-to-Speech (TTS) model CosyVoice2 [25]. Based on the effectiveness of the pre-trained speech model Wav2Vec2 [14] in dementia detection [26,27], we extract Wav2Vec2 embeddings from these synthesized speech and perform AD detection again. Remarkably, some ASR-synthesized speech, exhibits higher AD detection performance than the manual-synthesized speech. To our knowledge, no previous work has reported or explained this phenomenon, making it worthy of an in-depth investigation.

Understanding the decision-making process of deep learning-based medical systems is crucial for fostering transparency and trust in clinical applications. Iqbal et al. [28] used interpretability tools to identify key linguistic features in AD detection. Gimeno-Gómez et al. [29] developed a framework to uncover meaningful speech characteristics for Parkinson's disease detection. Inspired by these works, we propose a cross-attention-based interpretability model to explore why certain ASR transcripts (or synthesized speech) outperform manual ones in AD detection accuracy, and to elucidate the underlying mechanisms driving classification. Specifically, we first select some knowledge-based text and speech features (also used individually for AD detection), which are then fused with embeddings from BERT or Wav2Vec2 within our model. By analyzing attention scores from the best-performing models based on ASR and manual transcripts (or ASR- and manual-synthesized speech), we identify which knowledge-based features are enhanced by ASR errors, leading to their recognition as valuable cues for AD detection, ultimately improving classification accuracy. We also employ this model to unveil AD-related patterns withih pre-trained embeddings. The overall workflow is illustrated in Figure 1. Our main contributions are as follows:

- As the first comprehensive study to investigate AD detection using ASR transcripts with varying WER and their synthesized speech, advancing ASR-assisted AD detection.
- We find that certain ASR transcripts and ASR-synthesized speech outperform their manual counterparts, suggesting that ASR errors may provide valuable cues for AD detection.
- We propose a cross-attention-based interpretability model that identifies valuable cues from both transcript and speech perspectives, achieving superior or comparable performance and unveiling AD-related patterns within embeddings.

2. Dataset

We selected and organized three datasets from DB for fine-tuning the ASR models: WLS (187 samples), Lu (54 samples),

* Corresponding author.

10.21437/Interspeech.2025-1598

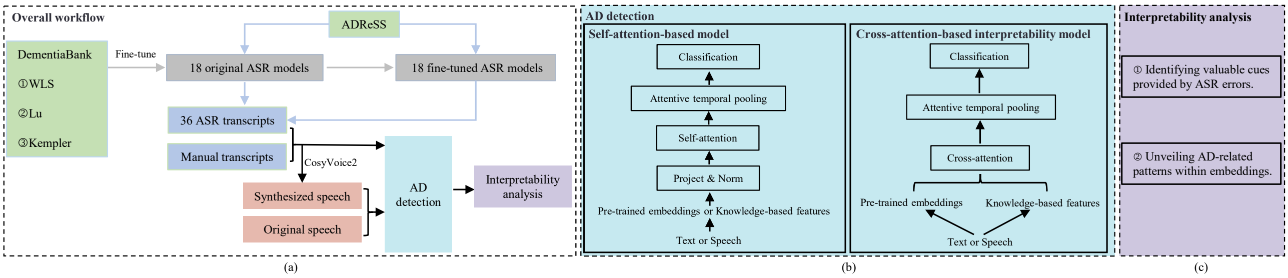


Figure 1: *(a) refers to the overall workflow, (b) refers to the AD detection models, and (c) refers to the interpretability analysis.*

and Kempler (4 samples). The ADReSS challenge dataset [8] from Interspeech 2020 was used for binary AD detection, with 108 training samples (54 AD, 54 HC) and 48 test samples (24 AD, 24 HC). Each sample consists of an audio recording and a corresponding manual transcript, which include subject's verbal descriptions of the "Cookie Theft" picture elicited by the interviewer. The manual transcripts were annotated in the CHAT format, which we processed to align with the subjects' actual speech. The WLS, Lu, and Kempler datasets were merged, resulting in 245 samples, which were split into ASR training (80%) and test sets (20%). Using timestamps from the CHAT annotations, we segmented the audio recordings into utterances that contained only the subjects' speech. We fine-tuned the ASR models on these utterance-level segments and transcripts. For ADReSS, ASR subject-level transcripts were created by concatenating the ASR utterance-level transcripts. The ASR subject-level transcripts (synthesized speech), manual transcripts (synthesized speech), and original speech of ADReSS were used for subsequent AD detection.

3. Methods

3.1. Fine-tuning ASR models

We selected 18 variants from Wav2Vec2 [14], HuBERT [15], WavLM [16], and Whisper [17] for fine-tuning, as these models provide varying WER. The selected variants are: ***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*}, ***whisper***-{*tiny, base, small, medium, large, large-v2, large-v3*}. All these ASR models are available on HuggingFace, and we fine-tuned them using the standard methods from the *Transformers* Python library.

3.2. Speech synthesis

To explore the impact of ASR transcripts with varying WER on AD detection from a speech perspective, we utilized CosyVoice2 [25] to synthesize speech from both ASR and manual transcripts. CosyVoice2 is an advanced TTS model that delivers nearly lossless synthesis. Using its zero-shot functionality, we randomly selected a segment from the subject's original speech as a prompt for each transcript, ensuring the synthesized speech closely matched the original characteristics.

3.3. Feature extraction

3.3.1. Knowledge-based features

To enable interpretable analysis, we extracted 35 and 60 knowledge-based features from each text and speech sample, respectively, chosen for their proven effectiveness as AD biomarkers in prior studies [28–30].

The 35 knowledge-based text features, extracted using the Python libraries *lexical-diversity*, *textstat*, and *spaCy*, as well as the MRC psycholinguistic database [31], were classified into six categories: **Lexical Diversity** (4): type-token ratio (TTR), mean segmental TTR (MSTTR), moving average TTR (MATTR), measure of lexical textual diversity (MTLD); **Content Complexity** (7): syllable count, stop words count, lexicon count, difficult words count, average sentence length, content density, propositional density; **Disfluencies** (2): the ratio of filler pauses (uh, um, er, and ah) and repetition words; **Part of Speech** (11): the ratio of Pronouns, Verbs, Nouns, Adjectives, Adverbs, Conjunctions, Articles, Determiners, Prepositions, Pronoun-Verb, and Pronoun-Noun; **Readability** (6): Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Coleman-Liau Index (CLI), Dale-Chall Readability Score (DCRS), Automated Readability Index (ARI); **Psycholinguistic Features** (5): familiarity, concreteness, imagability, meaningfulness, age of acquisition rating.

The 60 knowledge-based speech features, extracted using the DisVoice toolkit[1], were classified into four categories (the full description is in the toolkit above): **Articulation** (4): the mean and standard deviation (std) of the first (F1) and second (F2) formant frequencies; **Glottal** (14): the mean and std of seven descriptors: variability of time between consecutive glottal closure instants (GCI), average and variability of normalized amplitude quotient (NAQ) for consecutive glottal cycle, average and variability of opening quotient (OQ) for consecutive glottal cycles, average and variability of harmonic richness factor (HRF); **Phonation** (7): the mean of jitter, shimmer, amplitude perturbation quotient (APQ), pitch perturbation quotient (PPQ), and logarithmic energy (logE); the mean and std of the first derivative of the fundamental frequency (DF0); **Prosody** (35): the mean, std, maximum, minimum, skewness, and kurtosis of F0 contour; the mean, std, skewness, and kurtosis of energy contour for voice segments; number of voiced segments per second (NVSS); the mean, std, skewness, kurtosis, maximum, and minimum of voiced duration, unvoiced duration, and pause duration; the six duration ratios: pause/(voiced+unvoiced) (PVU), pause/unvoiced (PU), unvoiced/(voiced+unvoiced) (UVU), voiced/(voiced+unvoiced) (VVU), voiced/pause (VP), unvoiced/pause (UP).

3.3.2. Pre-trained embeddings

To capture rich characteristics for AD detection, we employed pre-trained BERT[2] and Wav2Vec2[3] models from HuggingFace to extract embeddings. Both models consisted of 24 transformer layers. Previous studies [26, 29, 32, 33] have shown that Wav2Vec2's intermediate layers capture more acoustic and phonetic information, while BERT's intermediate layers contain more syntactic information, offering superior AD detection performance compared to other layers. These findings align with our preliminary experiments. Therefore, we focused on the 8th layer of Wav2Vec2 and the 11th layer of BERT to better fuse these embeddings with knowledge-based features. For each text, we used BERT to generate embeddings of dimension $N \times 1024$, where N was the number of tokens (excluding

[1] https://github.com/jcvasquezc/DisVoice
[2] google-bert/bert-large-uncased
[3] jonatasgrosman/wav2vec2-large-xlsr-53-english

Table 1: *Mean WER (%) of ASR models on ADReSS dataset.*

ASR models	Original / Fine-tuned	ASR models	Original / Fine-tuned
w2v100	68.55 / 54.32	wavlm base	68.53 / 59.22
w2v960	61.05 / 50.80	wavlm large	57.76 / 54.13
w2v960 large	57.16 / 45.99	whisper tiny	62.58 / 43.33
w2v960 large lv	51.86 / 45.87	whisper base	58.10 / 36.39
w2v960 self	49.86 / 42.65	whisper small	49.91 / 30.47
w2v xlsr	55.67 / 42.19	whisper medium	45.70 / 29.39
w2v xlsr 1b	49.89 / 36.79	whisper large	45.32 / 28.04
hubert large	50.23 / 41.89	whisper large v2	45.98 / 28.65
hubert xlarge	49.18 / 45.99	whisper large v3	46.00 / **26.36**

[CLS] and [SEP]). For speech, we segmented each sample into 30-second chunks, extracted features with Wav2Vec2, and concatenated the segment-level embeddings along the time dimension to form subject-level embeddings of dimension $T \times 1024$, where T was the number of time steps.

3.4. AD detection models

The proposed AD detection models are illustrated in Figure 1.

3.4.1. Self-attention-based model

The model takes pre-trained embeddings or knowledge-based features as input. It includes a linear projection with layer normalization, a self-attention module, an attentive temporal pooling module [26, 34] for capturing richer temporal feature statistics, and a linear classification module. This model serves as a baseline for comparison with the cross-attention-based model.

3.4.2. Cross-attention-based interpretability model

The model utilizes knowledge-based features and pre-trained embeddings as inputs. According to Section 3.3, for each sample, we represented the knowledge-based feature as $X_{\text{kno}} \in \mathbb{R}^{1\times m}$ and the pre-trained embedding as $X_{\text{emb}} \in \mathbb{R}^{n\times 1024}$, where m was either 35 or 60, and n corresponded to the number of tokens or time steps. We modified the standard cross-attention mechanism to effectively integrate both representations while providing interpretability, as shown in Equation 1. Specifically, we first repeated X_{kno} n times to obtain $X'_{\text{kno}} \in \mathbb{R}^{n\times m}$. Then, we defined $Q = X_{\text{emb}} \cdot W_q$, $K = X'_{\text{kno}} \cdot W_k$, and $V = X_{\text{emb}} \cdot W_v$, where $W_q, W_v \in \mathbb{R}^{1024\times 1024}$ were learnable weight matrices, and W_k was an identity matrix I. As a result, the knowledge-based features directly served as K, providing a set of interpretable dimensions that were fused with the embeddings. d_k was the dimensionality of K, which was m. Next, the attention score matrix $A \in \mathbb{R}^{1024\times m}$ was computed via the scaled dot-product attention and softmax, capturing the relationship between the pre-trained embeddings and knowledge-based features. We utilized this matrix for subsequent interpretability analysis. Finally, the attention scores were applied to V, resulting in an enriched representation $Y \in \mathbb{R}^{n\times m}$, which was passed through the attentive temporal pooling module and a linear classifier to produce the final prediction.

$$Attention(Q, K, V) = V \cdot softmax\left(\frac{Q^T K}{\sqrt{d_k}}\right) \quad (1)$$

4. Experiments and results

4.1. Experimental setup

For fine-tuning the ASR models, we employed the AdamW optimizer with a learning rate of 1×10^{-5} and weight decay of 5×10^{-3}. The models were trained for 20 epochs with a batch size of 8, and performance was evaluated using WER. For AD detection, we used AdamW with a learning rate of 4×10^{-4} and weight decay of 1×10^{-5}. The training ran for 50 epochs with a batch size of 16, using cross-entropy loss. To ensure robustness, we conducted 10 trials with different random seeds and evaluated performance based on mean accuracy. All experiments were performed on NVIDIA RTX 4090 or A100 GPUs.

Table 2: *AD detection mean accuracy (%) on transcripts (manual + 36 ASR) and speech (original + synthesized from manual and 36 ASR transcripts). The three values separated by "/" represent the accuracy of three AD detection methods: from left to right, they correspond to inputting knowledge-based features / pre-trained embeddings into the self-attention model, and using the cross-attention model. In each AD detection method, the bold accuracy from ASR transcripts (ASR-synthesized speech) indicates its superiority over the corresponding manual transcripts (manual-synthesized speech).*

	Transcripts		Speech	
	Manual transcripts: 79.17 / 81.67 / 80.42		Original Speech: 75.83 / 77.92 / 78.75 Manual-synthesized: 59.58 / 74.58 / 70.42	
ASR models	Fine-tuned	Original	Fine-tuned	Original
w2v100	75.42 / 81.25 / 75.83	72.92 / 74.17 / 72.08	57.92 / 68.75 / 62.08	59.17 / 72.50 / 62.92
w2v960	75.83 / 77.92 / 74.17	76.25 / 76.67 / 76.25	**65.42** / 69.58 / 60.42	**66.67** / 67.08 / 65.83
w2v960 large	72.92 / 79.58 / 73.33	75.42 / 77.08 / 75.00	57.92 / 67.50 / 63.75	55.42 / 68.33 / 63.33
w2v960 large lv	76.67 / 76.67 / 77.50	69.58 / 74.17 / 72.08	55.42 / 65.83 / 64.17	55.00 / 67.08 / 59.17
w2v960 self	**81.25** / 80.83 / 79.58	75.83 / 77.92 / 75.42	**66.67** / 71.67 / 67.50	**69.58** / 70.83 / 64.17
w2v xlsr	74.17 / 75.83 / 73.33	72.50 / 77.08 / 69.38	**62.08** / 69.17 / 60.83	52.92 / 70.42 / 58.96
w2v xlsr 1b	78.33 / 75.42 / 77.08	77.92 / 79.58 / 74.58	**60.83** / 72.50 / 65.42	**62.08** / 68.75 / 62.08
hubert large	71.25 / **81.67** / 76.67	**81.25** / 78.75 / 74.38	59.17 / 72.08 / 63.33	**63.75** / 72.92 / 57.71
hubert xlarge	78.75 / 80.42 / 78.33	77.50 / 79.17 / 75.42	**62.08** / 70.00 / 65.83	**63.75** / 65.00 / 58.96
wavlm base	78.75 / 77.50 / 76.25	77.50 / 75.83 / 75.00	**63.75** / 64.17 / 59.58	**63.33** / 69.58 / 57.92
wavlm large	76.25 / **82.50** / **81.25**	75.42 / 75.42 / 76.25	**68.75** / 70.42 / 66.67	53.75 / 72.50 / 65.00
whisper tiny	77.50 / 78.75 / 79.17	**79.58** / 75.00 / 71.25	**62.08** / 72.08 / 67.50	**65.42** / 63.33 / 66.67
whisper base	**79.17** / 79.58 / **80.83**	72.50 / 76.67 / 72.92	**65.42** / 70.42 / 68.75	**67.50** / 67.50 / 64.17
whisper small	**80.00** / **82.50** / **81.25**	**80.42** / **82.08** / 76.67	**69.58** / **75.42** / **72.08**	**64.58** / 65.00 / 65.83
whisper medium	**80.42** / **82.92** / 78.33	**85.42** / **81.67** / 79.58	**62.92** / **75.83** / **71.67**	**60.00** / **75.00** / 67.08
whisper large	**79.58** / **83.33** / **82.50**	72.50 / **82.08** / 74.17	**62.08** / 68.33 / **70.83**	**67.50** / 68.75 / 65.42
whisper large v2	**85.42** / **84.58** / **80.42**	**80.83** / 77.92 / 77.08	**60.42** / 69.58 / **70.42**	**59.58** / **75.42** / 63.33
whisper large v3	77.08 / 79.17 / 80.00	77.08 / 78.33 / 75.83	**60.42** / 70.42 / **70.83**	**61.67** / **77.50** / 64.17

4.2. ASR models performance

Table 1 shows the mean WER of original and fine-tuned 18 ASR models on the ADReSS dataset. It can be observed that all ASR models achieved significant performance improvements after fine-tuning, with the fine-tuned *whisper large v3* achieving the lowest WER. Regardless of the WER values, the 36 ASR transcripts from ADReSS will be used in AD detection research.

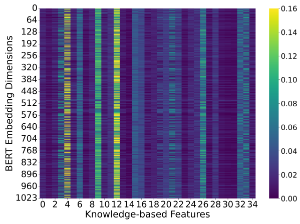
4.3. AD detection results

Table 2 presents the AD detection results on transcripts and speech. Key observations include: (a) In each detection method, certain ASR transcripts (ASR-synthesized speech) outperform manual transcripts (manual-synthesized speech) in accuracy, suggesting ASR errors may offer valuable cues for AD detection. We hypothesize that these errors introduce asymmetric biases between AD and HC groups, which provide useful cues that the model may exploit to improve accuracy. (b) The proposed cross-attention-based method outperforms or achieves comparable performance to the self-attention-based method on both original speech and manual transcripts, while also offering interpretability, highlighting its effectiveness and advantages. (c) Accuracy from synthesized speech is consistently lower than from original speech in all methods, indicating that the CosyVoice2 TTS model fails to fully capture pathological features in AD patients' speech. (d) Accuracy from manual transcripts exceeds that from original speech in all methods, suggesting linguistic features are more effective than acoustic features alone in AD detection. (e) In the self-attention-based method, pre-trained embeddings generally outperform knowledge-based features, showing that pre-trained embeddings capture richer, more comprehensive representations.

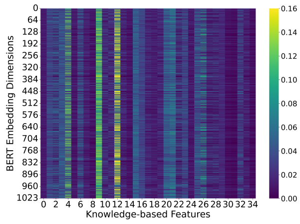
4.4. Cross-attention-based interpretability analysis

We explored the valuable cues provided by ASR errors from two perspectives: transcript and synthesized speech. Specifically, we compared and analyzed the attention scores generated by best-performing classification models on ASR and manual transcripts (or ASR- and manual-synthesized speech).

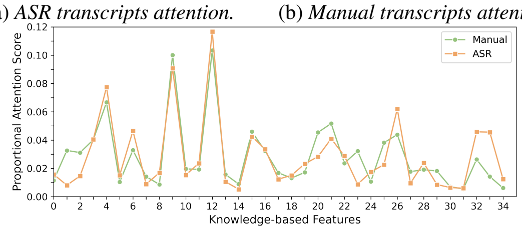
Transcript Perspective: We first selected the cross-attention-based model with the best-performing random seed on ASR transcripts (from fine-tuned *whisper large*, with the



(a) *ASR transcripts attention.*



(b) *Manual transcripts attention.*

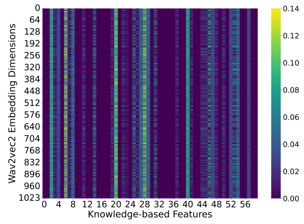


(c) *Attention comparison of ASR and manual transcripts.*

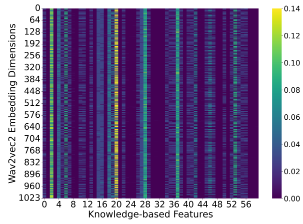
Figure 2: *Attention analysis of transcripts.*

best detection accuracy of 87.5%), then input the corresponding ADReSS test set into this model. For each correctly predicted sample, we obtained an attention score matrix (i.e., matrix A in Section 3.4.2), then averaged all matrices, as shown in Figure 2a. The x-axis (0-34) corresponded to each of the knowledge-based text features introduced sequentially in Section 3.3.1. This figure revealed the classification model's attention to different knowledge-based features and the relationship between pre-trained embeddings and knowledge-based features. A similar approach was applied to manual transcripts (the best detection accuracy was 83.33%), resulting in Figure 2b. We summed and normalized the two figures across the embedding dimension to obtain Figure 2c, providing a clearer contrast that highlighted which knowledge-based features received higher attention in ASR transcripts compared to manual transcripts. From Figure 2c, the primary features with higher attention in ASR transcripts include syllable count, lexicon count, average sentence length, filler pauses ratio, repetition words ratio, FRE, GFI, DCRS, concreteness, imagability, meaningfulness, and age of acquisition rating. This suggests that these features may show greater differences between the AD and HC groups in ASR transcripts, compared to manual transcripts, making them more discriminative and valuable cues for improving AD detection. Our additional statistical experiments support the hypothesis mentioned above: compared to manual transcripts, the AD group in ASR transcripts exhibits lower syllable count, lexicon count, and average sentence length, along with higher filler pauses and repetition words ratios. These ASR errors reflect the decline in language abilities of AD patients (e.g., more indistinct pronunciation and disfluencies) and amplify key linguistic features, indicating that ASR could serve not only as a transcription tool but also as an auxiliary tool for AD detection. Moreover, designing ASR models tailored for AD-related speech could enhance this effect.

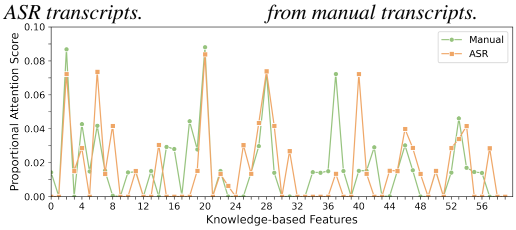
Synthesized Speech Perspective: We applied the same approach as in the transcript perspective, resulting in Figure 3. The x-axis (0–59) corresponded to the knowledge-based speech features introduced sequentially in Section 3.3.1. The best-performing model for ASR-synthesized speech, based on the fine-tuned *whisper small*, achieved 77.08% detection accuracy, while the best detection accuracy for manual-synthesized speech was 72.92%. From Figure 3c, the primary features with higher attention in ASR-synthesized speech compared to manual-synthesized speech include the mean, maximum, and skewness of pause duration; the maximum, minimum, and skewness of unvoiced duration; maximum of voiced duration; PVU; VVU; mean, maximum, minimum, and skewness of F0; std of F2; mean of NAQ and OQ variability; and the std of GCI variability and OQ average. Similar to the analysis from the transcript perspective, the features mentioned above may serve as valuable cues for AD detection. Since ASR errors can partially reflect the dysfluencies of AD patients, dysfluency-related features in ASR-synthesized speech may play a more significant role, which aligns with the features related to pause, voiced, and unvoiced duration listed above.



(a) *Attention of speech synthesized from ASR transcripts.*



(b) *Attention of speech synthesized from manual transcripts.*



(c) *Attention comparison of speech synthesized from ASR and manual transcripts.*

Figure 3: *Attention analysis of synthesized speech.*

4.5. AD-relevant patterns within pre-trained embeddings

We employed attention scores to unveil the knowledge-based features encoded in the pre-trained embeddings from both manual transcripts and original speech (focusing on intrinsic characteristics rather than comparative analysis with ASR). For manual transcripts, Figures 2b and 2c (green line) show that the repetition words ratio (12), content density (9), and syllable count (4) exhibit higher attention or correlation with the BERT embedding. For original speech, a similar attention score heatmap and line plot (omitted due to space limitations) reveal that PVU, along with the maximum and minimum voiced duration, exhibit higher attention or correlation with the Wav2Vec2 embeddings. These findings highlight the significance of these features in AD detection. Additionally, we found that almost all knowledge-based features are not encoded into specific dimensions of the pre-trained embeddings, but instead emerge through complex interrelationships across nearly all dimensions, suggesting that almost all embedding dimensions contribute significantly.

5. Conclusions

In this paper, we conducted a comprehensive study on AD detection using ASR transcripts with various WER and their corresponding synthesized speech. Our results indicated that ASR errors may offer valuable cues for improving AD detection, which we attributed to the asymmetric biases introduced by these errors between the AD and HC groups. Additionally, we proposed a cross-attention-based interpretability model that not only identified valuable cues from both transcript and speech perspectives but also achieved superior or comparable performance to the baseline model. We further leveraged this model to unveil AD-related patterns within pre-trained embeddings. This study provided new insights into the potential of ASR models in AD detection. Our future work will focus on designing ASR models tailored for AD speech to maximize this potential.

6. Acknowledgements

This work was partially supported by the Anhui Province Major Science and Technology Research Project (Grant No. S2023Z20004), by the National Social Science Foundation of China (Grant No. 23AYY012), and by the Supercomputing Center of the University of Science and Technology of China.

7. References

[1] P. J. Nestor, P. Scheltens, and J. R. Hodges, "Advances in the early detection of alzheimer's disease," *Nature medicine*, vol. 10, no. Suppl 7, pp. S34–S41, 2004.

[2] T. Talkar, S. Charles, C. Krantsevich, and K. Kawabata, "Detection of cognitive impairment and alzheimer's disease using a speech- and language-based protocol," in *Proc. Interspeech*, 2024, pp. 3025–3029.

[3] Y.-L. Liu, R. Feng, J.-H. Yuan, and Z.-H. Ling, "Clever hans effect found in automatic detection of alzheimer's disease through speech," in *Proc. Interspeech*, 2024, pp. 2435–2439.

[4] C. Botelho, J. Mendonça, A. Pompili, T. Schultz, A. Abad, and I. Trancoso, "Macro-descriptors for alzheimer's disease detection using large language models," in *Proc. Interspeech*, 2024, pp. 1975–1979.

[5] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden *et al.*, "Using the outputs of different automatic speech recognition paradigms for acoustic-and BERT-based Alzheimer's dementia detection through spontaneous speech." in *Proc. Interspeech*, 2021, pp. 3810–3814.

[6] K. Mei, X. Ding, Y. Liu, Z. Guo, F. Xu, X. Li, T. Naren, J. Yuan, and Z. Ling, "The USTC system for ADReSS-M challenge," in *Proc. ICASSP*, 2023, pp. 1–2.

[7] Y.-L. Liu, R. Feng, J.-H. Yuan, and Z.-H. Ling, "Leveraging prompt learning and pause encoding for alzheimer's disease detection," in *Proc. ISCSLP*, 2024, pp. 486–490.

[8] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," in *Proc. Interspeech*, 2020, pp. 2172–2176.

[9] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," *arXiv preprint arXiv:2104.09356*, 2021.

[10] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual Alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge," in *Proc. ICASSP*, 2023, pp. 1–2.

[11] S. D. L. F. Garcia, F. Haider, D. Fromm, B. MacWhinney, A. Lanzi, Y.-N. Chang *et al.*, "Connected speech-based cognitive assessment in chinese and english," *arXiv preprint arXiv:2406.10272*, 2024.

[12] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for Alzheimer's dementia through spontaneous speech." in *Proc. Interspeech*, 2020, pp. 2222–2226.

[13] J. Li, J. Yu, Z. Ye, S. Wong, M. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for Alzheimer's disease detection," in *Proc. ICASSP*, 2021, pp. 6423–6427.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, pp. 12 449–12 460.

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[18] L. Zhou, K. C. Fraser, F. Rudzicz *et al.*, "Speech recognition in Alzheimer's disease and in its assessment." in *Proc. Interspeech*, 2016, pp. 1948–1952.

[19] A. Balagopalan, K. Shkaruta, and J. Novikova, "Impact of ASR on Alzheimer's disease detection: All errors are equal, but deletions are more equal than others," *arXiv preprint arXiv:1904.01684*, 2019.

[20] C. Li, W. Xu, T. Cohen, and S. Pakhomov, "Useful blunders: Can automated speech recognition errors improve downstream dementia classification?" *Journal of Biomedical Informatics*, vol. 150, p. 104598, 2024.

[21] Y.-L. Liu, R. Feng, Y.-X. Lu, J.-X. Chen, Y. Ai, J.-H. Yuan, and Z.-H. Ling, "Can automated speech recognition errors provide valuable clues for alzheimer's disease detection?" in *Proc. ICASSP*, 2025, pp. 1–5.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[23] Y. Wang, T. Wang, Z. Ye, L. Meng, S. Hu, X. Wu, X. Liu, and H. Meng, "Exploring linguistic feature and model combination for speech recognition based automatic ad detection," in *Proc. Interspeech*, 2022, pp. 3328–3332.

[24] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease." in *Proc. Interspeech*, 2020, pp. 2162–2166.

[25] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[26] J. Li, K. Song, J. Li, B. Zheng, D. Li, X. Wu, X. Liu, and H. Meng, "Leveraging pretrained representations with task-related keywords for alzheimer's disease detection," in *Proc. ICASSP*, 2023, pp. 1–5.

[27] Y. Pan, B. Mirheidari, D. Blackburn, and H. Christensen, "A two-step attention-based feature combination cross-attention system for speech-based dementia detection," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[28] F. Iqbal, Z. S. Syed, M. S. S. Syed, and A. S. Syed, "An explainable ai approach to speech-based alzheimer's dementia screening," in *Proc. SMM*, 2024, pp. 11–15.

[29] D. Gimeno-Gómez, C. Botelho, A. Pompili, A. Abad, and C.-D. Martínez-Hinarejos, "Unveiling interpretability in self-supervised speech representations for parkinson's diagnosis," *arXiv preprint arXiv:2412.02006*, 2024.

[30] P. Pérez-Toro, S. Bayerl, T. Arias-Vergara, J. Vásquez-Correa, P. Klumpp, M. Schuster, E. Nöth, J. Orozco-Arroyave, and K. Riedhammer, "Influence of the interviewer on the automatic assessment of alzheimer's disease in the context of the adresso challenge," in *Proc. Interspeech*, 2021, pp. 3785–3789.

[31] M. Wilson, "Mrc psycholinguistic database: Machine-usable dictionary, version 2.00," *Behavior research methods, instruments, & computers*, vol. 20, no. 1, pp. 6–10, 1988.

[32] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in *Proc. ACL*, 2019.

[33] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ASRU*, 2021, pp. 914–921.

[34] C. d. Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," *arXiv preprint arXiv:1602.03609*, 2016.