Event-Duration Distribution Modeling and Co-Adapted Ensembles for Speech-Based ADRD Screening

Wei Bo, Manav Kanaganapalli, Shuwei Hou, Xiaoyu Zhang, Anarghya Das, and Wenyao Xu
Department of Computer Science and Engineering
University at Buffalo, State University of New York
Buffalo, NY, USA
weibo, manavkan, shuweiho, zhang376, anarghya, wenyaoxu@buffalo.edu

***Abstract*—Speech-based screening for Alzheimer's disease and related dementias (ADRD) offers a non-invasive approach to early detection, but impaired speech patterns arise from heterogeneous mechanisms that are not adequately captured by single-modality representations. Temporal disruption poses a distinct encoding challenge, as conventional approaches either reduce it to scalar statistics or force it into sequential models, neither of which preserves the severity structure. This paper proposes CAEEDD (Co-Adaptive Event-Distribution-Driven Detection), a modality- and structure-aware framework that integrates symbolic linguistic features, continuous acoustic dynamics, and event-distribution representations capturing the severity, variability, and tail behavior of temporal disruption, and employs feature–model co-adaptation to align heterogeneous feature subsets with classifiers of matched inductive biases. On the Pitt corpus, CAEEDD achieves 88.41% accuracy, outperforming scalar temporal baselines (60.87%), sequence models (up to 81.88%), and zero- and few-shot LLMs (up to 69.13%), with a mean improvement of 16.81% over scalar baselines sustained across five random stratified splits. An ablation analysis further demonstrates the complementary contributions of linguistic, acoustic, and temporal cues. These findings suggest that distribution-aware multimodal representations and feature–model co-adaptation provide a more effective strategy for speech-based ADRD screening and motivate further evaluation across diverse tasks, populations, and clinical settings.**

***Index Terms*—ADRD screening, multimodal representation, temporal disruption, event-duration distribution, feature-model co-adaptation.**

I. Introduction

Alzheimer's Disease and Related Dementias (ADRD) represent a rapidly growing public health challenge, with more than 6.9 million affected adults in the United States alone, with annual economic costs exceeding 360 billion dollars [1]. These burdens disproportionately affect underserved populations where access to neuroimaging and invasive biomarker assays is limited, underscoring the need for scalable, non-invasive digital biomarkers for early detection and longitudinal monitoring of cognitive decline.

Speech has emerged as a promising modality for this purpose, as it engages multiple cognitive systems, including lexical retrieval, working memory, executive control, and motor planning, that are progressively disrupted across the ADRD continuum. Acoustic, linguistic, and temporal characteristics of connected speech can distinguish cognitively impaired individuals from healthy controls, often preceding overt clinical symptoms [2], [3], and speech can be collected repeatedly and remotely using commodity devices, making it well suited for large-scale screening [4]. Related work has shown that daily-life speech can serve as a low-burden digital biomarker for disease screening [5].

Despite this promise, a fundamental methodological challenge remains in how speech is represented for clinical modeling. Speech-based ADRD analysis involves heterogeneous phenomena, including temporal disruption, lexical access, and discourse organization, that differ in modality, statistical structure, and temporal scale [6]. Acoustic signals capture continuous prosodic and articulatory variability, transcript-derived features encode discrete symbolic processes, and temporal fluency indicators behave as irregular event-duration processes [7]–[9]. Projecting such heterogeneous signals into a single homogeneous representation risks suppressing modality-specific structure and reducing complementary diagnostic information [10]. This challenge is widely recognized in multimodal machine learning, where effective systems preserve modality-specific representations while enabling integration across heterogeneous sources [11].

Within this broader challenge, temporal disruption, manifested as abnormal pause behavior, prolonged hesitations, and irregular timing, illustrates the core encoding difficulty. Existing approaches largely follow two paradigms. Aggregate methods reduce timing behavior to summary statistics such as pause counts or duration measures, which are interpretable but often discard severity structure, variability, and rare long-duration events [7], [12], [13]. Order-sensitive models instead treat speech as a sequential signal and learn temporal dependencies using recurrent or neural sequence architectures, but such assumptions may be unstable when hesitation events are irregular, speaker-dependent, and weakly coupled in small clinical corpora [10], [14], [15]. Recent findings suggest that pause behavior in cognitive impairment is better characterized

by distributional properties, including dispersion, multimodality, and heavy-tailed patterns, rather than by global averages or temporal order alone [8], [16].

These representation issues are further amplified by the small, noisy, and speaker-variable nature of clinical speech datasets, where performance is sensitive to feature selection and model inductive bias [4], [6], [10]. A single classifier may not exploit all feature families equally well, because linguistic, acoustic, and temporal features differ in sparsity, scale, and statistical structure. Ensemble learning provides a principled way to combine complementary predictors, and stacked generalization allows different model families to specialize on distinct feature subsets while producing a unified decision [17], [18]. These observations motivate feature-model co-adaptation, in which feature subsets are aligned with classifiers that have compatible inductive biases.

Motivated by this insight, we propose a framework that aligns modeling strategies with both the structure of individual speech phenomena and the constraints of low-resource clinical data. Our contributions are threefold. First, we propose CAEEDD (Co-Adaptive Event-Distribution-Driven Detection), a modality- and structure-aware framework with a feature-model co-adaptation ensemble for speech-based ADRD screening. CAEEDD integrates symbolic linguistic indicators, continuous acoustic dynamics, and event-distribution representations of temporal disruption, while aligning heterogeneous feature subsets with classifiers of matched inductive biases. Second, we introduce Event-Duration Distributions as a representation of temporal disruption that captures severity structure, variability, and tail behavior beyond conventional scalar pause descriptors while avoiding assumptions of stable sequential dependency. Third, we evaluate CAEEDD on the Pitt Cookie Theft corpus [2] against scalar temporal baselines, order-sensitive sequence models, transcript-based large language models, and representation-group ablations, providing empirical insight into how complementary speech cues contribute under limited-data clinical conditions.

II. Methodology

A. Overview

We propose CAEEDD, a modality- and structure-aware framework for speech-based ADRD screening. As shown in Fig. 1, the framework takes filtered subject audio recordings and aligned CHAT-format transcripts as input, extracts three complementary representation groups, including continuous acoustic dynamics, event-distribution temporal disruption, and symbolic linguistic indicators, and integrates them through a feature–model co-adaptation ensemble. Instead of collapsing heterogeneous speech cues into a single uniform representation, CAEEDD preserves modality-specific feature structure and allows different learners to independently operate on feature subsets aligned with their inductive biases before decision-level aggregation. The framework is evaluated through comparative experiments across multiple modeling paradigms and use representation-group ablations to quantify the contribution of each feature family under small clinical dataset constraints.

B. Dataset, Preprocessing, and Evaluation Protocol

Experiments are conducted on the Pitt Corpus (DementiaBank) [2] using Cookie Theft picture-description recordings. We use both audio recordings and corresponding CHAT-format transcripts, which provide speaker-attributed utterances and annotations for linguistic and interactional analysis.

To prevent speaker leakage, we use a stratified 75/25 split at the participant level, ensuring that all recordings from the same participant remain in the same partition. The primary comparative evaluation is conducted on a fixed stratified split for reproducibility, with all models evaluated on the same held-out test partition. The test partition is used only for final evaluation; feature normalization, feature selection, model selection, and ensemble construction are performed within the training partition, as detailed in Section II-E.

To assess split sensitivity, we repeated the primary CAEEDD configuration and key representation-restricted configurations across five stratified participant-level splits. These configurations included scalar pause descriptors only, event-distribution temporal features only, continuous acoustic features only, symbolic linguistic features only, and symbolic linguistic plus acoustic features. The ensemble structure and per-model feature counts were fixed to the primary-split configuration, while feature selection, model fitting, and stacked aggregation were repeated within each training partition. Sequence models and LLM baselines were not repeated across splits. We report mean and standard deviation as a secondary stability analysis to complement the fixed-split results.

Audio recordings and CHA transcripts are temporally aligned to obtain word/utterance boundaries, speaker turns, and silence intervals. We define temporal disruptions as silence intervals exceeding 250 ms, which serve as the basis for deriving distributional pause statistics and discrete event encodings.

C. Event-Duration Distribution Modeling of Temporal Disruption

We model temporal disruption as an irregular event-duration process derived from aligned audio-transcript data. Let

$$\mathcal{E} = \{(t_i, d_i)\}_{i=1}^{N}$$

denote the set of disruption events, where t_i is the onset time and d_i is the duration of the i-th silence interval exceeding $\Delta = 250$ ms. To preserve duration severity without imposing a continuous sequential model, each disruption is quantized as

$$k_i = \left\lfloor \frac{d_i}{\Delta} \right\rfloor,$$

where k_i represents the number of minimum-duration units in the disruption. The empirical distribution of $\{k_i\}$ captures *micro-level* temporal severity, including frequent short hesitations and rare prolonged pauses. For *meso-level* fragmentation, disruption events partition speech into production bursts:

$$\mathcal{B} = \{b_1, b_2, \ldots, b_M\},$$

where b_j denotes the duration of a continuous speech segment between consecutive disruptions. For *macro-level* structure,

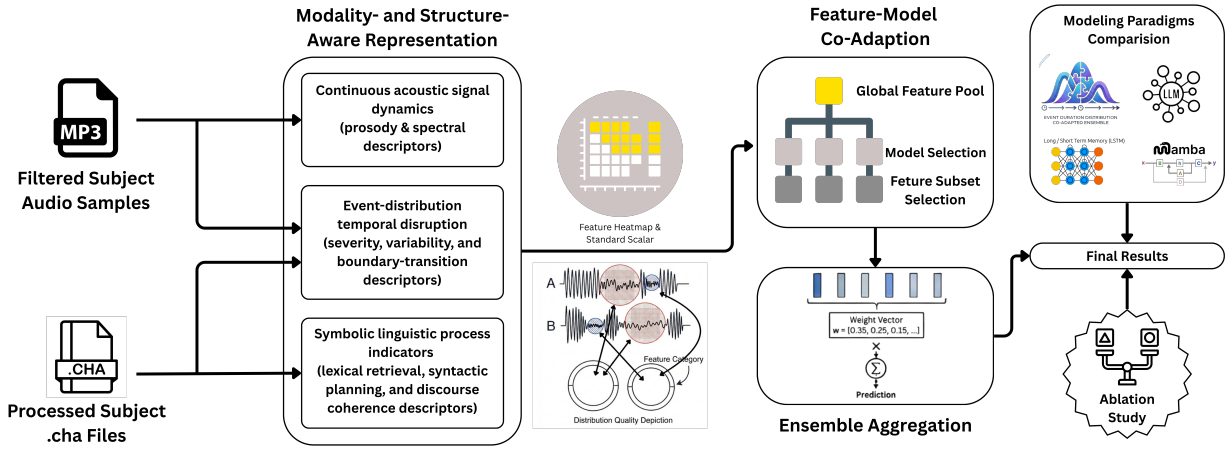

Fig. 1. Overview of the CAEEDD framework. Multimodal speech features extracted from audio and transcripts are organized into modality- and structure-aware representation groups, followed by feature–model co-adaptation and ensemble aggregation for ADRD detection.

boundary annotations and speaker turns from CHA transcripts define discourse segments

$$\mathcal{S} = \{s_1, s_2, \ldots, s_K\}.$$

Together, temporal representation is defined as

$$\mathbf{x}^{(T)} = \Phi(\mathcal{E}, \mathcal{B}, \mathcal{S}),$$

where $\Phi(\cdot)$ extracts distributional and structural descriptors across micro-, meso-, and macro-level dynamics. This representation preserves duration variability, speech-burst fragmentation, and discourse-level segmentation while avoiding assumptions of stable long-range sequential dependency.

D. Modality- and Structure-Aware Representation

To operationalize CAEEDD, we organize 107 features by input modality and statistical structure of the underlying speech phenomena.

Symbolic linguistic representations (58 features) are derived from CHAT transcripts and capture lexical diversity, repetition/perseveration, incomplete utterances, discourse segmentation, CHAT markers such as filled pauses and retracing events, and transcript-level grammatical and semantic patterns [9], [19]. These features provide a strong transcript-derived representation of language organization and content.

Continuous acoustic representations (18 features) are extracted from audio signals and capture prosodic and spectral dynamics, including pitch and energy variability, speaking-time and acoustic pause statistics, MFCC variability, spectral dispersion, and spectral flux [20], [21]. These features preserve continuous speech-production cues that are not fully represented in transcripts.

Event-based temporal representations (31 features) are derived from the event-duration modeling in Section II-C by combining audio timing with transcript-based structural markers [7], [8], [22]. They include pause-distribution descriptors, speech-burst/continuity measures, and boundary/transition features that characterize temporal disruption across hesitation, fragmentation, and discourse structure.

This grouping treats acoustic and temporal descriptors as complementary to the strong linguistic representation and enables the subsequent co-adaptation framework to match heterogeneous feature subsets with classifiers of appropriate inductive biases.

E. Feature-Model Co-Adaptation Ensemble

Because the feature groups differ in statistical structure, CAEEDD uses a feature-model co-adaptation ensemble that matches each base learner with a model-specific feature subset and combines predictions through stacked ensemble learning.

Candidate learners include linear, kernel-based, and tree-based ensemble models. For each learner, feature subsets are selected using feature importance ranking or recursive feature elimination, and candidate subset sizes are evaluated by stratified 5-fold cross-validation on the training partition. The configuration with the highest mean validation accuracy is retained. To avoid optimistic bias, all feature selection, subset-size optimization, base learner selection, and ensemble construction are performed within the training partition only. The test partition is not consulted and is used only for final evaluation. Final ensemble composition is selected from retained base learners using out-of-fold stacking on the training partition, and the resulting out-of-fold predictions are used to train the meta-learner.

This design allows different classifiers to specialize in compatible feature subsets while combining complementary decision patterns, providing a bias-aware strategy for integrating heterogeneous speech representations under limited-data constraints.

F. Comparative Evaluation Across Modeling Paradigms

We evaluate CAEEDD against representative baselines that reflect three alternative assumptions about ADRD-related speech impairment: scalar temporal summarization, order-sensitive temporal modeling, and transcript-centered language reasoning.

Classical temporal baselines use coarse scalar pause and segment statistics, such as pause counts, pause durations, and segment counts. This comparison tests whether event-duration distributions capture temporal disruption beyond global pause summaries.

Order-sensitive baselines include a bidirectional LSTM and a Mamba-based state-space model trained on ordered event-level representations under the same train/test partition and optimization protocol. This comparison examines whether preserving event order improves classification relative to distribution-based temporal modeling under limited-data conditions.

Large language models (LLMs) are evaluated as transcript-centered reference classifiers using representative GPT-, Claude-, Gemini-, and Llama-family models. Each model receives the participant's transcript augmented with summary pause statistics and is prompted with a fixed template to output a binary dementia/control label in zero-shot or few-shot settings. This comparison tests whether off-the-shelf language reasoning over transcripts and coarse timing summaries can substitute for task-specific multimodal representations. Because these models do not receive acoustic signals or the full engineered feature set, their results are interpreted within this transcript-centered evaluation scope rather than as fully modality-matched multimodal baselines.

III. Experiments and Results

A. Experimental Setup

Experiments were conducted on the Pitt Corpus Cookie Theft task [2] using a stratified 75/25 participant-level split to preserve class balance and prevent speaker leakage. Performance was evaluated using accuracy and weighted F1-score, with all models assessed on identical partitions. Unless otherwise specified, results correspond to the primary split at seed 789; repeated-split results are reported separately as a stability analysis.

B. Results of Feature-Model Co-Adaptation Ensemble

Following the procedure described in Section II-E, we evaluated multiple base learners, including Support Vector Classifier (SVC) [23], Random Forest [24], Extremely Randomized Trees (ExtraTrees) [25], Linear Discriminant Analysis (LDA) [26], Gradient Boosting Machine (GBM) [27], and Extreme Gradient Boosting (XGBoost) [28]. Each learner was paired with a model-specific feature subset selected through feature-importance ranking or Recursive Feature Elimination [29]. Candidate subset sizes and learner combinations were evaluated through out-of-fold stacking within the training partition.

As shown in Table I, ExtraTrees achieved the highest individual accuracy (84.78%), followed by GBM (83.33%), while SVC and Random Forest each achieved 81.88%. LDA and XGBoost achieved 76.81%. The variation across learners indicates that predictive performance depends on the interaction between the classifier and its selected feature representation, supporting model-specific feature adaptation. The final stacked ensemble combined SVC, LDA, and GBM, integrating kernel-based, linear discriminant, and tree-boosting decision mechanisms. It achieved 88.41% accuracy and a weighted F1-score of 0.8841, exceeding the strongest individual learner. The ensemble correctly classified 53 of 61 control samples and 69 of 77 dementia samples, supporting the benefit of combining complementary co-adapted learners under limited-data conditions.

TABLE I
Performance of individual base learners and CAEEDD.

Model	Accuracy (%)	Weighted F1
SVC (25 features)	81.88	0.82
Random Forest (70 features)	81.88	0.82
ExtraTrees (35 features)	84.78	0.85
LDA (20 features)	76.81	0.77
Gradient Boosting (30 features)	83.33	0.83
XGBoost (30 features)	76.81	0.77
CAEEDD (proposed)	**88.41**	**0.8841**

C. Comparative Evaluation Results

Table II compares CAEEDD with a scalar pause baseline, order-sensitive sequence models, and transcript-based large language models (LLMs). The proposed CAEEDD achieved the highest performance, with 88.41% accuracy and a weighted F1-score of 0.88. The scalar pause-count baseline achieved 60.87% accuracy, indicating that coarse aggregate descriptors do not preserve the variability, severity structure, and tail behavior captured by the event-duration representation. This interpretation is consistent with prior evidence that cognitive impairment is associated with increased variability and heavier-tailed pause distributions [8], [22].

Sequence models, including bidirectional long short-term memory network (Bi-LSTM) and Mamba state-space model, achieved 81.88% and 78.99% accuracy, respectively, but did not surpass CAEEDD. A likely reason is that hesitation events in spontaneous ADRD speech occur irregularly and vary substantially across speakers, providing limited stable ordering patterns for sequence models to learn from a small clinical dataset. Under these conditions, long-range sequence models may capture speaker-specific variation or noise, whereas the event-duration representation preserves severity, dispersion, and tail behavior without requiring consistent event ordering. These results suggest that distribution-aware representations offer a more stable characterization of temporal disruption under the present data constraints.

Transcript-based LLMs achieved 64.33-69.13% accuracy under zero- and few-shot prompting, with few-shot Claude Sonnet performing best at 69.13%. Each model received the Cookie Theft transcript and structured pause summaries and produced a binary label under a fixed prompt, but had no access to acoustic signals, fine-grained temporal features, or task-specific training. Although few-shot demonstrations improved classification consistency, transcript-centered inference may still overlook variability and severity patterns embedded in the original speech signal. This asymmetric comparison

TABLE II
COMPARATIVE PERFORMANCE ACROSS MODELING PARADIGMS

Approach	Accuracy (%)	Weighted F1
CAEEDD (Proposed)	88.41	0.88
Scalar Baseline (Pause Counts)	60.87	0.62
Bi-LSTM	81.88	0.80
Mamba (State-Space Model)	78.99	0.75
Claude Sonnet (Few-shot)	69.13	0.72
Claude Haiku (Few-shot)	66.54	0.69
Gemini Flash (Few-shot)	66.48	0.70
Llama 3.3 70B (Zero-shot)	65.99	0.66
GPT-4o-mini (Zero-shot)	64.33	0.64

is intended to characterize the gap between text-accessible and multimodal representations rather than establish modality-matched parity.

D. Repeated-Split Stability Analysis

To evaluate whether the primary findings depended on a favorable train–test partition, we repeated the analysis across five stratified participant-level splits (seeds 789, 42, 123, 456, 999) using the fixed ensemble configuration. CAEEDD achieved the highest mean accuracy at 79.13% (±4.79%), exceeding all representation-restricted configurations: Symbolic Linguistic Process Only at 78.41% (±2.98%), the combined symbolic linguistic and continuous acoustic configuration at 77.10% (±3.68%), Event-Distribution Temporal Disruption Only at 67.25% (±4.86%), Continuous Acoustic Dynamics Only at 64.93% (±2.27%), and Scalar Pause Descriptors Only at 62.32% (±4.85%). CAEEDD consistently outperformed the scalar baseline across all five splits, with a mean margin of 16.81%, demonstrating that the benefit over coarse temporal summaries was not driven by a single favorable partition. The smaller mean difference from the linguistic-only configuration reflects the strong task-specific value of high-quality transcript features; the complementary contribution of acoustic and temporal features is examined further through the paired error analysis in the ablation study.

E. Analysis of Pause Severity Structure

To examine the information captured by the distribution-aware temporal representation, pauses were grouped into five duration bins: very short (≤ 0.5 s), short (0.5-1.5 s), medium (1.5-3 s), long (3-5 s), and very long (> 5 s), following the modeling procedure described in Section II-C. These bins distinguish brief pauses associated with routine speech planning from increasingly prolonged hesitations that may reflect greater difficulty in lexical retrieval and discourse production.

Figure 2 presents the normalized importance of the five bins across the SVC, LDA, and Gradient Boosting components. Short pauses contributed consistently across models, whereas long and very long pauses showed greater model dependence, suggesting that rare prolonged events provide complementary rather than uniformly discriminative information. Figure 3 further shows higher pause counts in the dementia group

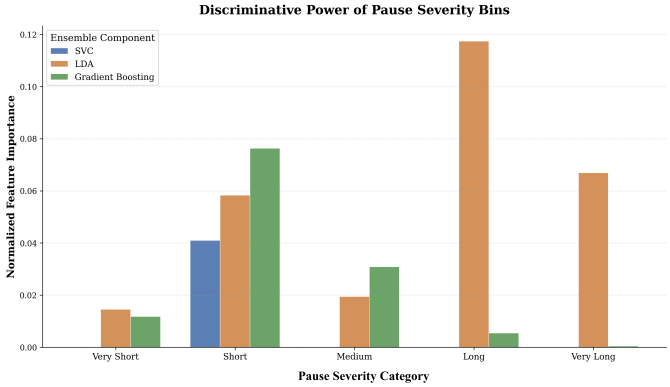

Fig. 2. Discriminative power of pause severity bins across ensemble components. Feature importance values indicate the contribution of each severity category to model decisions.

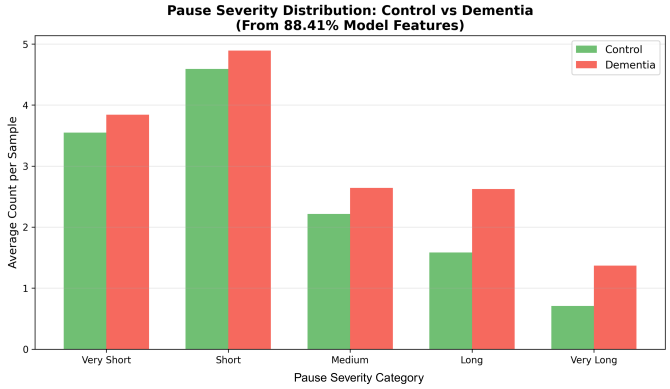

Fig. 3. Average number of pauses per severity bin for control and dementia participants. Dementia speech shows increased counts in longer pause bins, indicating heavier-tailed temporal disruption.

across all bins, with the largest differences for long and very long pauses. While short pauses may reflect normal planning or clause boundaries, the heavier-tailed profile in dementia speech suggests greater variability and more frequent prolonged disruptions associated with lexical retrieval and maintaining the flow of thought [7], [22]. These findings support modeling both the frequency and severity distribution of pauses rather than aggregate counts alone.

F. Ablation Study of Representation Groups

To examine the contribution of the representation groups defined in Section II-D, we evaluated CAEEDD using restricted subsets of the modality- and structure-aware features (Table III). The Event-Distribution Temporal Disruption Only configuration achieved 66.67% accuracy, compared with 60.87% for Scalar Pause Descriptors Only, indicating that preserving the severity and variability of pause durations provides information beyond aggregate counts.

The Symbolic Linguistic Process Only configuration achieved 84.78%, reflecting the strong discriminative value of high-quality CHAT transcripts in the Cookie Theft task [9]. Nevertheless, the full CAEEDD model improved accuracy to

TABLE III
ABLATION STUDY ON REPRESENTATION GROUPS

Configuration	Accuracy (%)	Change (%)
CAEEDD (Proposed)	88.41	–
Scalar Pause Descriptors Only	60.87	-27.54
Event-Distribution Temporal Disruption Only	66.67	-21.74
Continuous Acoustic Dynamics Only	65.22	-23.19
Symbolic Linguistic Process Only	84.78	-3.62
Symbolic Linguistic + Continuous Acoustic	86.23	-2.18

88.41% and corrected six errors made by the linguistic-only model while introducing only one new error, including two recovered dementia false negatives and four corrected control false positives. Although McNemar's test did not reach conventional significance ($\chi^2 = 3.57$, $p = 0.059$), the paired error pattern provides directional evidence that acoustic and temporal representations contribute complementary information and improve the balance between sensitivity and specificity. For context, transcript-based LLMs receiving the same transcripts and structured pause summaries achieved only 64.33-69.13%, suggesting that the strong linguistic-only result is attributable to the engineered lexical and discourse representations rather than transcript text alone.

The Symbolic Linguistic and Continuous Acoustic configuration achieved 86.23%, improving over linguistic features alone but remaining below the full framework. This result suggests that simple multimodal combination captures some complementary information, whereas feature-model co-adaptation more effectively aligns heterogeneous representations with model-specific inductive biases. Overall, the ablation results support the complementary value of acoustic and temporal cues beyond transcript-derived features.

IV. DISCUSSION

A. Insights Revealed by Modeling Heterogeneous Speech Impairments

The findings support the view that ADRD-related speech abnormalities arise from partially distinct mechanisms, including linguistic planning deficits, motor/prosodic impairment, and production-timing disruption [3], [4]. Their relative importance therefore depends on the task and individual presentation. In the Cookie Theft benchmark, transcript-derived features were highly discriminative, while acoustic and temporal representations provided complementary information that corrected errors not resolved by linguistic features alone. This pattern favors multimodal integration without implying that every modality contributes equally across settings.

The results also suggest that irregular hesitation events are not optimally characterized by either aggregate pause counts or order-sensitive sequence models in this case. Pause behavior varied substantially across and within speakers, limiting the stability of learnable event-order patterns in a small clinical corpus. Event-duration distributions instead preserved severity, dispersion, and extreme-event behavior without requiring consistent temporal ordering, supporting the interpretation of disruption as an irregular event process [8], [16].

Finally, performance differences across learners and representation subsets indicate that heterogeneous features interact differently with model inductive biases. Allowing classifiers to specialize on compatible feature subsets and combining their predictions at the decision level reduced the limitations of uniform feature fusion. More broadly, these findings suggest that speech-based ADRD screening benefits from representation-aware integration that respects the distinct statistical structure of linguistic, acoustic, and temporal phenomena.

B. Clinical Implications and Practical Considerations

The findings suggest that speech-based ADRD screening may benefit from capturing multiple dimensions of impairment rather than relying on a single modality. Linguistic features were highly informative in the structured Cookie Theft task, while acoustic and temporal representations captured complementary motor and production-timing cues. These signals may be especially useful in settings involving spontaneous speech, variable recording conditions, or imperfect transcription, where transcript quality can be affected by automatic speech recognition errors, language background, education, and disease stage. Integrating acoustic and temporal information could therefore reduce overreliance on transcript-derived indicators, although this benefit should be tested directly using degraded or ASR-generated transcripts [4], [6]. Prior multimodal ADRD work likewise showed benefits from complementary information [30].

The paired error analysis further showed that CAEEDD recovered two dementia cases missed by the linguistic-only model while correcting four false positives. This balance is clinically relevant because false negatives may delay assessment and intervention, whereas false positives can lead to unnecessary follow-up. Multimodal integration may therefore improve the reliability of screening decisions beyond changes in aggregate accuracy alone. Evaluation across additional elicitation tasks, populations, and recording environments is still needed before broader clinical generalizability can be established.

C. Limitations and Future Directions

This study is limited to a single English-language structured task and a relatively small clinical corpus. The importance of linguistic, acoustic, and temporal representations may differ across conversational settings, languages, populations, and disease stages. Moreover, the modest gain over the linguistic-only configuration indicates that the benefit of multimodal integration is task- and data-dependent and should be examined directly under degraded or ASR-generated transcripts.

The current comparison does not include pre-trained self-supervised speech representations such as wav2vec 2.0, HuBERT, and WavLM. CAEEDD instead emphasizes interpretable representations and learning under the data constraints

of small clinical corpora. Future evaluation will examine whether event-distribution features provide complementary information to pre-trained speech embeddings.

The implementation also relies partly on CHAT-format transcripts with speaker and interactional annotations. Although the acoustic features and most event-based temporal features are derived from aligned audio, some symbolic linguistic features require expert annotation. Future work will prioritize evaluation across additional speech tasks and the development of an automated ASR- and diarization-based pipeline.

V. Conclusion

This paper presents CAEEDD, a modality- and structure-aware framework for speech-based ADRD screening that integrates linguistic, acoustic, and event-based temporal representations through feature-model co-adaptation. Results on the Pitt Corpus show that distribution-aware temporal features preserve severity and variability patterns that are lost in coarse pause summaries, while the co-adapted ensemble effectively integrates heterogeneous representations under limited-data conditions. Although transcript-derived features were highly discriminative, paired error analysis showed that acoustic and temporal cues provided complementary information beyond linguistic features alone. Overall, the findings demonstrate the value of representation-aware multimodal integration within the evaluated setting and motivate further validation across additional speech tasks and datasets.

References

[1] A. Association, "Alzheimer's facts and figures report — alzheimer's association," 2025. [Online]. Available: https://www.alz.org/alzheimers-dementia/facts-figures

[2] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," Archives of neurology, vol. 51, no. 6, pp. 585–594, 1994.

[3] I. Vigo, L. Coelho, and S. Reis, "Speech-and language-based classification of alzheimer's disease: a systematic review," Bioengineering, vol. 9, no. 1, p. 27, 2022.

[4] J. Robin, J. E. Harrison, L. D. Kaufman, F. Rudzicz, W. Simpson, and M. Yancheva, "Evaluation of speech-based digital biomarkers: review and recommendations," Digital biomarkers, vol. 4, no. 3, pp. 99–108, 2020.

[5] S. Bhattacharjee and W. Xu, "VoiceLens: A multi-view multi-class disease classification model through daily-life speech data," Smart Health, vol. 23, p. 100233, 2022.

[6] X. Qi, Q. Zhou, J. Dong, and W. Bao, "Noninvasive automatic detection of alzheimer's disease from spontaneous speech: a review," Frontiers in Aging Neuroscience, vol. 15, p. 1224723, 2023.

[7] J. Yuan, X. Cai, Y. Bian, Z. Ye, and K. Church, "Pauses for detection of alzheimer's disease," Frontiers in Computer Science, vol. 2, p. 624488, 2021.

[8] P. Pastoriza-Dominguez, I. G. Torre, F. Dieguez-Vide, I. Gómez-Ruiz, S. Geladó, J. Bello-López, A. Ávila-Rivera, J. A. Matias-Guiu, V. Pytel, and A. Hernández-Fernández, "Speech pause distribution as an early marker for alzheimer's disease," Speech Communication, vol. 136, pp. 107–117, 2022.

[9] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," IEEE transactions on audio, speech, and language processing, vol. 19, no. 7, pp. 2081–2090, 2011.

[10] K. Ding, M. Chetty, A. Noori Hoshyar, T. Bhattacharya, and B. Klein, "Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges," Artificial Intelligence Review, vol. 57, no. 12, p. 325, 2024.

[11] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 2, pp. 423–443, 2018.

[12] J. Liu, F. Fu, L. Li, J. Yu, D. Zhong, S. Zhu, Y. Zhou, B. Liu, and J. Li, "Efficient pause extraction and encode strategy for alzheimer's disease detection using only acoustic features from spontaneous speech," Brain Sciences, vol. 13, no. 3, p. 477, 2023.

[13] R. A. Sluis, D. Angus, J. Wiles, A. Back, T. Gibson, J. Liddle, P. Worthy, D. Copland, and A. J. Angwin, "An automated approach to examining pausing in the speech of people with dementia," American Journal of Alzheimer's Disease & Other Dementias®, vol. 35, p. 1533317520939773, 2020.

[14] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech," arXiv preprint arXiv:2106.09668, 2021.

[15] A. Roshanzamir, H. Aghajan, and M. Soleymani Baghshah, "Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech," BMC Medical Informatics and Decision Making, vol. 21, no. 1, p. 92, 2021.

[16] D. He, L. Xu, T. Betthauser, C. Van Hulle, J. Liss, V. Berisha, and K. Mueller, "Automated bimodal pause analysis for acoustic markers of cognitive decline and alzheimer's disease in connected speech," Alzheimer's & Dementia, vol. 21, no. 9, p. e70635, 2025.

[17] T. G. Dietterich, "Ensemble methods in machine learning," in International workshop on multiple classifier systems. Springer, 2000, pp. 1–15.

[18] D. H. Wolpert, "Stacked generalization," Neural networks, vol. 5, no. 2, pp. 241–259, 1992.

[19] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," Journal of Alzheimer's disease, vol. 49, no. 2, pp. 407–422, 2015.

[20] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert et al., "Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease," Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 1, no. 1, pp. 112–124, 2015.

[21] I. Hajjar, M. Okafor, J. D. Choi, E. Moore, A. Abrol, V. D. Calhoun, and F. C. Goldstein, "Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early alzheimer's disease," Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 15, no. 1, p. e12393, 2023.

[22] M. Lofgren and W. Hinzen, "Breaking the flow of thought: increase of empty pauses in the connected speech of people with mild and moderate alzheimer's disease," Journal of Communication Disorders, vol. 97, p. 106214, 2022.

[23] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[24] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[25] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3–42, 2006.

[26] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, vol. 7, no. 2, pp. 179–188, 1936.

[27] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[29] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine Learning, vol. 46, no. 1, pp. 389–422, 2002.

[30] W. Bo, S. S. Sullivan, X. Zhang, M. Gao, and W. Xu, "A telemedicine analytic framework for fully and semi-automatic alzheimer's disease screening using clock drawing test," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 12, pp. 7503–7516, 2024.