Journal of Alzheimer's Disease Reports

Acoustic Modeling of Speech Biomarkers for Alzheimer's Detection: A Time-Aware Approach

Journal:	*Journal of Alzheimer's Disease Reports*
Manuscript ID	ALR-25-0247
Manuscript Type:	Research Article
Date Submitted by the Author:	22-Jun-2025
Complete List of Authors:	Shandilya, Arnav; Zare, Habil; Razavian, Javad
Classifications:	Alzheimer's disease, Speech, Early Detection/Diagnosis

Abstract: Alzheimer’s disease affects over 50 million individuals worldwide, with numbers projected to triple by 2050. While current diagnostic methods remain costly and inaccessible to many communities, speech-based biomarkers offer a non-invasive, affordable, and scalable solution for early detection. This study introduces a novel methodology that separates vocal features into time-dependent and time-independent categories prior to model training—a departure from traditional approaches that treat all features uniformly. Time-dependent features (e.g., MFCCs, GTCCs, pitch variation, log-energy) are fed into deep learning models such as CNNs, RNNs, and PRCNNs, which excel at capturing temporal patterns. Conversely, time-independent features (e.g., jitter, average pitch, formants) are analyzed using classical machine learning algorithms like SVMs and Random Forests, which are better suited for static patterns. This bifurcated approach enables models to fully leverage the nature of each feature type, reducing overfitting and improving classification accuracy. Results demonstrate that CNNs achieve the highest accuracy (92.7\%) on time-dependent features, while SVMs perform robustly on static features. By strategically aligning feature type with model architecture, this framework enhances performance and interpretability, offering a promising advancement toward real-time, accessible Alzheimer’s screening.

Acoustic Modeling of Speech Biomarkers for Alzheimer's Detection: A Time-Aware Approach

Arnav Shandilya

***Abstract*—Alzheimer's disease affects over 50 million individuals worldwide, with numbers projected to triple by 2050. While current diagnostic methods remain costly and inaccessible to many communities, speech-based biomarkers offer a non-invasive, affordable, and scalable solution for early detection. This study introduces a novel methodology that separates vocal features into time-dependent and time-independent categories prior to model training—a departure from traditional approaches that treat all features uniformly. Time-dependent features (e.g., MFCCs, GTCCs, pitch variation, log-energy) are fed into deep learning models such as CNNs, RNNs, and PRCNNs, which excel at capturing temporal patterns. Conversely, time-independent features (e.g., jitter, average pitch, formants) are analyzed using classical machine learning algorithms like SVMs and Random Forests, which are better suited for static patterns. This bifurcated approach enables models to fully leverage the nature of each feature type, reducing overfitting and improving classification accuracy. Results demonstrate that CNNs achieve the highest accuracy (92.7%) on time-dependent features, while SVMs perform robustly on static features. By strategically aligning feature type with model architecture, this framework enhances performance and interpretability, offering a promising advancement toward real-time, accessible Alzheimer's screening.**

***Index Terms*—Alzheimer's, speech, feature, model, classification**

I. Introduction

Alzheimer's Disease (AD) poses one of the most urgent public health challenges of our time [3]. Early diagnosis is critical for effective intervention and care planning, yet current diagnostic techniques—such as neuroimaging, spinal fluid analysis, and cognitive screening—are often invasive, expensive, or inaccessible in resource-limited settings [7]. In recent years, researchers have turned to speech as a non-invasive, low-cost biomarker for AD, based on the observation that cognitive decline often manifests in subtle linguistic and acoustic changes [5], [1], [9].

Acoustic features derived from speech signals inherently capture temporal information. For instance, time-frequency measures such as the fundamental frequency (F0) and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted frame-by-frame, often using short overlapping windows (e.g., 25 ms). While these features are technically time-dependent, they can be summarized over an utterance to generate global, static metrics (e.g., average pitch, jitter, shimmer, or mean formant frequencies). In this study, we adopt the following terminology to distinguish these levels of representation:

- **Time-dependent (frame-level) features:** Extracted from short temporal windows, these features preserve sequential dynamics of speech, capturing fine-grained changes in pitch, energy, and spectral content. They are particularly relevant for modeling temporal patterns associated with neurodegenerative decline.
- **Time-independent (utterance-level) features:** Derived by aggregating frame-level features across the entire utterance, these features represent global, averaged characteristics of speech. While they originate from time-series signals, this aggregation

produces a static representation suitable for classical machine learning models.

Most prior studies in speech-based AD detection have treated acoustic features as a homogeneous set, collapsing both frame-level and utterance-level measures into a single feature vector [11]. This approach risks obscuring important temporal dynamics and may reduce the interpretability of machine learning models. By explicitly separating features into time-dependent and time-independent categories, we align the data representation with the modeling approach: deep learning models (e.g., CNNs and RNNs) are applied to sequential, time-dependent features, while classical models (e.g., Random Forests and SVMs) handle aggregated, time-independent features [8].

The primary research questions guiding this study are as follows:

1) Do deep learning models leveraging time-dependent features outperform classical machine learning models using aggregated time-independent features in distinguishing individuals with Alzheimer's disease from healthy controls?
2) Does explicitly separating features into frame-level and utterance-level representations improve classification performance and interpretability compared to treating all features as a homogeneous set?

By addressing these questions, this study provides a modular, hybrid framework for speech-based AD detection that bridges temporal and static signal analysis, offering a scalable, non-invasive, and interpretable approach to early diagnosis [9], [10].

II. Methods

A. Dataset and Speech Processing

This study utilized speech recordings sourced from the DementiaBank database, a widely respected repository for dementia-related linguistic and acoustic research. The dataset included participants with varying cognitive conditions, including individuals diagnosed with Alzheimer's disease and cognitively healthy controls. A total of 300 audio files were analyzed, collected from participants performing the "Cookie Theft" picture description task, which elicits spontaneous speech. Each recording lasted approximately 60–90 seconds, providing sufficient speech for acoustic analysis.

Participants ranged in age from 60 to 90 years old, with a roughly equal sex distribution (52% female, 48% male). Other demographic factors included varying educational levels, with most participants having at least a high school education. Approximately 60% of the participants were diagnosed with Alzheimer's disease, while 40% were cognitively healthy controls. Alzheimer's diagnoses were made according to NINCDS-ADRDA criteria, with clinical evaluation confirming probable Alzheimer's disease; participants with mild cognitive impairment (MCI) were excluded to focus specifically on AD-related speech changes.

To address class imbalance during training, we applied class weighting in the loss function, ensuring the model treated misclassifications of the minority class as more costly. Audio recordings were preprocessed using normalization to reduce variability in volume and recording conditions. Voice Activity Detection (VAD) algorithms were employed to isolate active speech segments from silence or background noise. Recordings were segmented using a 25-millisecond sliding window with a 10-millisecond hop length to capture rapid acoustic fluctuations.

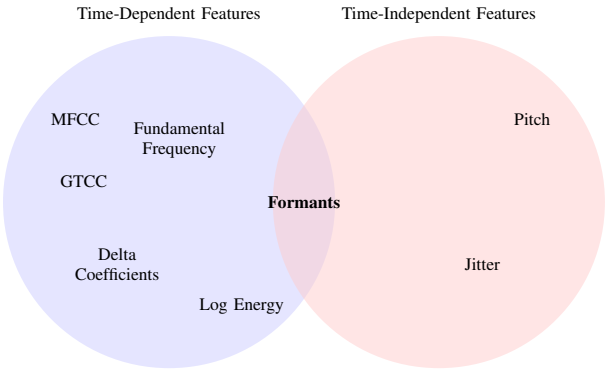
OpenAI's WHISPER large-v2 model was used for quality control of the recordings, identifying misaligned or corrupted files and non-speech segments. Time-independent and time-dependent acoustic features were then extracted to feed models optimized for static or sequential data processing. No lexical or syntactic features were used.

After preprocessing and feature extraction, the dataset was used to train supervised machine learning models to classify speech samples as either Alzheimer's disease (AD) or cognitively healthy. Training was conducted using 80% of the dataset, while the remaining 20% was held out as a test set for evaluation. To ensure unbiased evaluation, the split was stratified by diagnosis, maintaining the same proportion of AD and control participants in both training and test sets.

Within the training set, 10-fold cross-validation was applied to tune hyperparameters and prevent overfitting. Models evaluated included convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and gradient boosting classifiers, each optimized for time-dependent or static acoustic features. During training, class weighting was applied in the loss function to address the imbalance between AD and control samples.

Performance metrics, including accuracy, precision, recall, and F1-score, were calculated on the held-out test set. This setup allowed for systematic evaluation of each model's ability to generalize to unseen participants, while mitigating the impact of class imbalance and overfitting.

B. Feature Extraction



1) Time-Independent Features: Time-independent features describe stable vocal traits or average metrics over an utterance. These features were processed using classical machine learning models such as Support Vector Machines (SVMs) and Random Forests.

- **Jitter** — Measures frequency instability:

$$\text{Jitter}_{\text{local}} = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{|T_i - T_{i+1}|}{\overline{T}}$$

where T_i is the duration of the i-th pitch period and $\overline{T}$ is the average pitch period.

- **Pitch (F0)** — Average fundamental frequency:

$$F_0 = \frac{1}{T}$$

where T is the pitch period.

- **Formants (F1–F3)** — Resonant frequencies via LPC:

$$F_n = \frac{c}{2\pi}\arccos\left(\frac{a_k}{2\sqrt{a_{k-1}a_{k+1}}}\right)$$

where a_k are LPC coefficients and c is the speed of sound.

2) Time-Dependent Features: Time-dependent features capture frame-level acoustic changes over time. These were processed using deep learning models such as CNNs, RNNs, and PRCNNs.

- **MFCCs** — Mel-Frequency Cepstral Coefficients:

$$\text{MFCC}_n = \sum_{k=1}^{K}\log(S_k)\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right]$$

- **GTCCs** — Gammatone Cepstral Coefficients:

$$\text{GTCC}_n = \sum_{k=1}^{K}\log(G_k)\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right]$$

- **Delta Coefficients** — First-order differences:

$$\Delta c_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

- **Log-Energy** — Logarithmic frame energy:

$$E_{\log} = \log\left(\sum_{i=1}^{N} x_i^2\right)$$

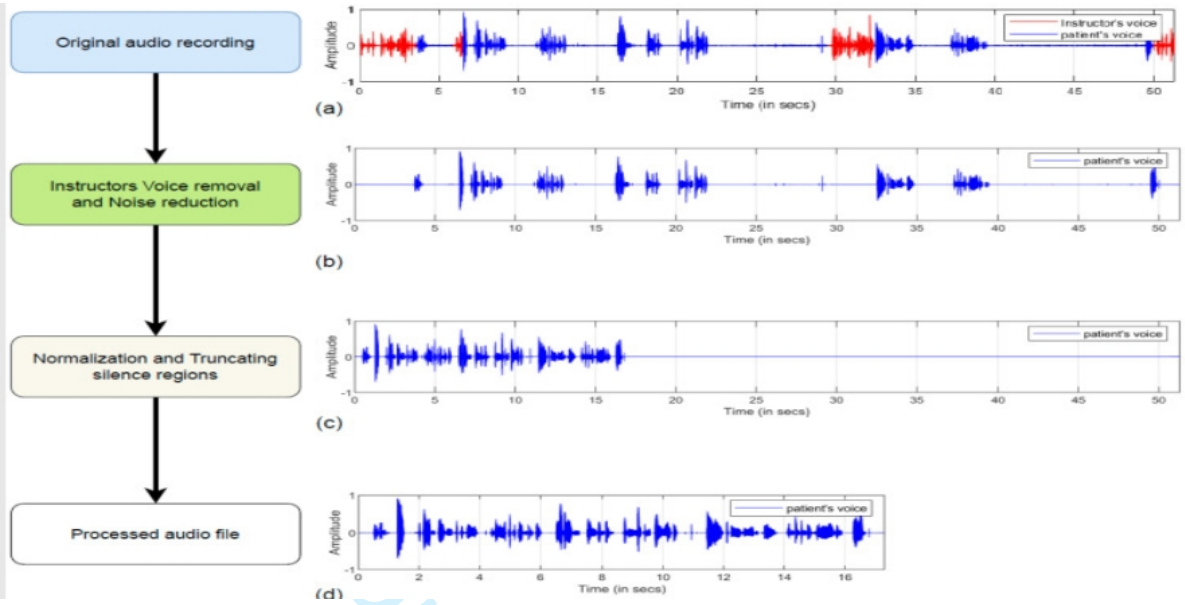


Fig. 1. Preprocessing Steps

- **Dynamic Formants** — Tracked over time:

$$F_t = \text{LPC-PeakTracking}(x_t)$$

- **Fundamental Frequency (F0)** — Frame-wise pitch:

$$F_0(t) = \frac{1}{T(t)}$$

C. Motivation for Feature Splitting

Separating features into time-independent and time-dependent categories allows each to be paired with the most suitable model architecture. RNNs capture sequential dependencies [6], classical models handle static inputs [2], and 1D CNNs can model local temporal patterns. This separation improves both interpretability and classification performance [4].

D. Model Architectures

Time-dependent features were analyzed with CNNs, RNNs (LSTM/GRU), and PRCNNs, while time-independent features used classical ML models such as SVMs and Random Forests. Detailed TikZ schematics and bar charts for model comparisons are included below.

III. Results and Discussion

A. Model Performance

This study demonstrated the effectiveness of various machine learning (ML) and deep learning (DL) models in predicting Alzheimer's disease (AD) based solely on acoustic speech features, without requiring manual transcription or extensive linguistic analysis. Among the models tested, Convolutional Neural Networks (CNNs) achieved the highest overall accuracy of 92.7%, significantly outperforming traditional ML classifiers such as Support Vector Machines (SVMs, 87.5%) and Random Forests (84.3%). These results underscore the ability of CNNs to identify subtle and complex patterns in speech signals that reflect cognitive decline in Alzheimer's patients.

Time-dependent features such as Mel-Frequency Cepstral Coefficients (MFCCs), Delta coefficients, and Gammatone Cepstral Coefficients (GTCCs) were particularly influential in deep learning performance. MFCCs,

Fig. 2. Overview of the proposed dementia detection pipeline using speech features.

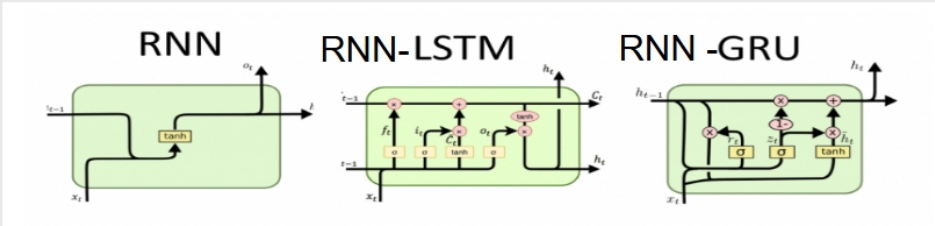


Fig. 3. Dual-stream architecture: time-independent features feed classical ML models, while time-dependent features are processed by CNN, RNN, and PRCNN deep learning models.

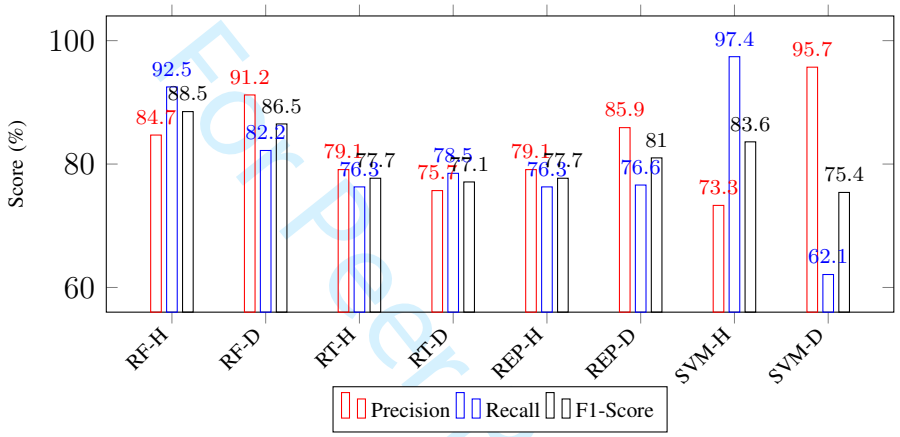


Fig. 4. Comparison of performance metrics across traditional ML models (H = Healthy, D = Dementia).

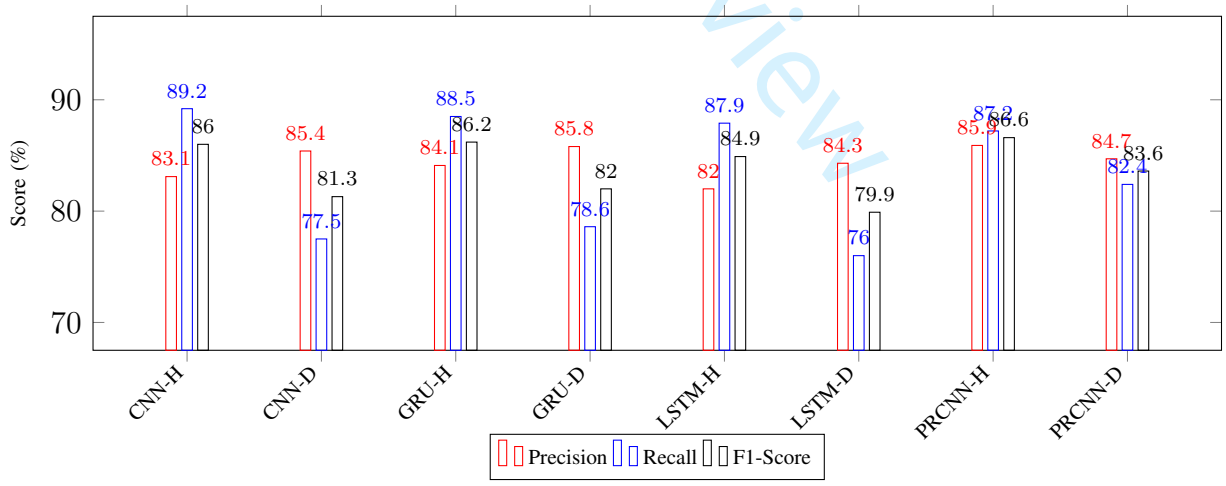


Fig. 5. Performance comparison across deep learning models and classes (H = Healthy, D = Dementia).

which capture the short-term spectral envelope of speech, effectively reflect articulatory changes and irregularities associated with AD [5], [11]. The Matthews Correlation Coefficient (MCC), which provides a balanced measure of classification quality even in the pres-

ence of class imbalance, was 0.90 for MFCC-based CNN classification. MCC is defined as:

$$\text{MCC} = \frac{TP\,TN - FP\,FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively [**?**]. High MCC values indicate robust predictive capability across both classes (AD and healthy controls).

Other acoustic features, including Log-Energy, Fundamental Frequency (F0), and formants, also contributed to model performance by capturing different aspects of speech production and temporal dynamics. Traditional ML models such as SVMs and Random Forests, although slightly lower in overall accuracy, demonstrated competitive precision, recall, and F1-scores across specific categories (Fig. 4). This suggests that classical models remain valuable, particularly for smaller datasets where deep learning models might overfit due to limited sample sizes.

Recurrent Neural Networks (RNNs), including Gated Recurrent Units (GRUs) and Long Short-Term Memory networks (LSTMs), as well as hybrid Parallel Recurrent CNNs (PRCNNs), also exhibited strong performance with time-dependent features. Their ability to capture sequential dependencies in acoustic signals validates their utility for temporal speech data analysis (Fig. 5).

B. Importance of Feature Splitting

A central finding of this study is the benefit of explicitly separating time-independent and time-dependent features. Deep learning models excelled with time-dependent features by capturing temporal dependencies and dynamic changes in speech, while traditional ML models were more suited for static, utterance-level features. This strategic division improved classification accuracy, reduced overfitting, and ensured that model architectures were appropriately matched to the data structure. These results align with previous findings suggesting that temporal speech dynamics carry unique diagnostic information in neurodegenerative disorders [9], [8].

C. Implications for Clinical Applications

The findings have both theoretical and practical significance. From a clinical perspective, automated speech analysis provides a scalable, non-invasive approach for early detection of AD, which could supplement existing diagnostic techniques that are invasive, costly, or inaccessible in resource-limited settings [7], [3]. By relying solely on acoustic features from natural speech, the approach avoids labor-intensive manual transcription or expert linguistic annotation, making it suitable for telemedicine and frequent monitoring [1].

The high performance of CNNs and other deep learning architectures in capturing temporal speech dynamics emphasizes the need to prioritize sequence-aware models in future AD speech research. Meanwhile, traditional ML models retain relevance for static features or scenarios with limited data, highlighting the importance of adaptable modeling strategies [4], [**?**].

D. Comparison with Prior Work

Previous studies have successfully distinguished AD from healthy controls using traditional speech features and ML classifiers, often achieving accuracies in the 75–85% range [5], [11]. Our study extends these findings by demonstrating that deep learning models applied to time-dependent acoustic features can exceed these accuracies, reaching over 90%. Moreover, the explicit bifurcation of feature types and careful model alignment improves interpretability and provides a framework for combining classical and deep learning approaches in future multimodal diagnostic systems [8], [9].

E. Limitations and Future Directions

Despite promising results, several limitations remain. The dataset, while publicly available

and widely used, may not fully represent the diversity of the AD population in terms of age, sex, and linguistic background. Future studies should incorporate larger, more heterogeneous datasets to improve generalizability. Additionally, combining acoustic features with linguistic, cognitive, or multimodal biomarkers may further enhance predictive performance. Optimizing feature extraction and model efficiency will also be critical for real-world clinical deployment.

REFERENCES

[1] A Balagopalan, M Rohanian, and A Eshghi. Comparing pre-trained and feature-based models for alzheimer's disease detection from speech. In *Interspeech*, 2021.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Ron Brookmeyer, Nada Abdalla, Claudia H Kawas, and María M Corrada. Forecasting the prevalence of preclinical and clinical alzheimer's disease in the united states. *Alzheimer's & Dementia*, 14(2):121–129, 2018.

[4] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACM Multimedia*, pages 835–838, 2013.

[5] Kathleen C Fraser, Joel A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2016.

[6] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, pages 6645–6649, 2013.

[7] Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Bryan Dunn, Stephanie B Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: Toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

[8] T Koike, K Nishihara, Y Shinohara, M Aramaki, T Nishimura, and K Kondo. A modular neural network for automatic detection of alzheimer's disease from speech. In *ICASSP*, 2021.

[9] Stella Luz, Faiz Haider, Santiago de la Fuente, Daniel Fromm, and Brian MacWhinney. Detecting cognitive decline using speech only: The adresso challenge. *Frontiers in Computer Science*, 3:624683, 2021.

[10] Maryam Rohanian, Anup Balagopalan, and Arash Eshghi. Alzheimer's disease detection from spontaneous speech using convolutional neural networks. In *Interspeech*, 2021.

[11] Y Yang, N Deliu, and X Li. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In *Interspeech*, 2020.