SpeechT-RAG: Reliable Depression Detection in LLMs with Retrieval-Augmented Generation Using Speech Timing Information

Xiangyu Zhang[1], **Hexin Liu**[2*], **Qiquan Zhang**[1*], **Beena Ahmed**[1], **Julien Epps**[1]
The University of New South Wales[1], Nanyang Technological University[2]

Abstract

Large Language Models (LLMs) have been increasingly adopted for health-related tasks, yet their performance in depression detection remains limited when relying solely on text input. While Retrieval-Augmented Generation (RAG) typically enhances LLM capabilities, our experiments indicate that traditional text-based RAG systems struggle to significantly improve depression detection accuracy. This challenge stems partly from the rich depression-relevant information encoded in acoustic speech patterns — information that current text-only approaches fail to capture effectively. To address this limitation, we conduct a systematic analysis of temporal speech patterns, comparing healthy individuals with those experiencing depression. Based on our findings, we introduce Speech Timing-based Retrieval-Augmented Generation, SpeechT-RAG, a novel system that leverages speech timing features for both accurate depression detection and reliable confidence estimation. This integrated approach not only outperforms traditional text-based RAG systems in detection accuracy but also enhances uncertainty quantification through a confidence scoring mechanism that naturally extends from the same temporal features. Our unified framework achieves comparable results to fine-tuned LLMs without additional training while simultaneously addressing the fundamental requirements for both accuracy and trustworthiness in mental health assessment.

1 Introduction

Large Language Models (LLMs)(Brown et al., 2020; Achiam et al., 2023) have been extensively utilized in health-related applications, achieving notable success in tasks such as medical evidence summarization(Tang et al., 2023), supporting decision-making in general surgery (Oh et al., 2023), and assisting in gastroenterological diagnoses (Lahat et al., 2023). These models have demonstrated exceptional capabilities in transforming traditional healthcare workflows by efficiently processing complex medical data and enhancing decision-making processes. However, in specialized domains like depression detection, LLMs based solely on text have shown limited effectiveness (Ohse et al., 2024; Zhang et al., 2024d; Wu et al., 2023b), often requiring extensive task-specific fine-tuning and large annotated datasets to achieve satisfactory results. This reliance on significant training resources makes their application to depression detection both time-consuming and costly.

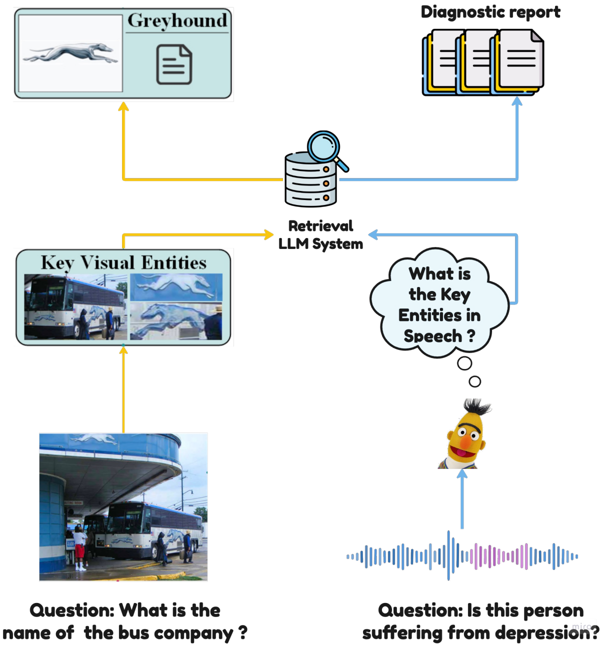

Greyhound
Diagnostic report
Retrieval
LLM System
Key Visual Entities
What is the Key Entities in Speech ?
Question: What is the name of the bus company ?
Question: Is this person suffering from depression?

Figure 1: Motivation for Speech Timing-based RAG. While visual RAG systems can leverage distinct entities as retrieval keys (left), identifying analogous "key entities" in speech for depression detection is challenging.

When the performance of LLMs is restricted, Retrieval-Augmented Generation (RAG) is often employed as an alternative. By allowing LLMs to retrieve task-specific information from external sources during inference, RAG has shown promise in mitigating the need for extensive fine-tuning

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 10019–10030
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

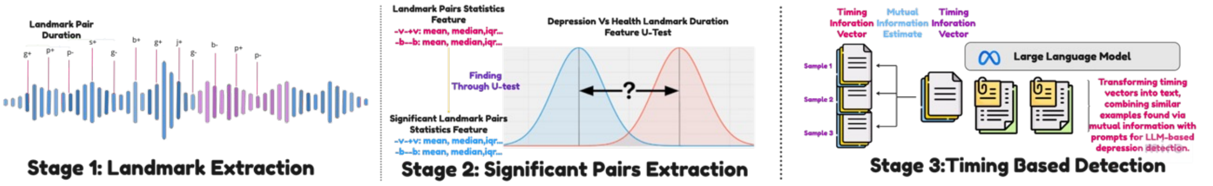


Figure 2: Overview of method: (1) extract acoustic landmarks and encode temporal information using durations between consecutive bigram pairs; (2) identify statistically significant landmark pairs that differentiate depression from health based on duration features; and (3) utilize timing keys and text-based representations for RAG.

across many tasks (Guu et al., 2020; Lewis et al., 2020). However, in the context of depression detection, the limited performance of RAG highlights the inherent constraints of text-based approaches shown in our experiment. These limitations stem from text's inability to capture critical non-verbal cues, such as tone, prosody, and timing, which play a significant role in identifying depression (Williamson et al., 2013; Quatieri and Malyska, 2012). Therefore, it is crucial to address these gaps by seeking additional sources of information that can complement text-based methods and overcome their inherent limitations.

Numerous studies have demonstrated that acoustic features in speech, particularly temporal and prosodic patterns, contain rich information highly relevant to depression detection, and many existing works have successfully leveraged these acoustic characteristics as the primary modality for this task (Huang et al., 2019a; Zhao et al., 2020). However, directly integrating acoustic features into text-based LLMs has often been detrimental to their performance (Zhang et al., 2024a,d), making multimodal RAG a promising alternative. In computer vision, RAG systems frequently utilize salient visual elements or regions of interest (such as objects or scene segments) as input to retrieve relevant information (Jian et al., 2024; Xia et al., 2024), but an analogous technique for identifying and utilizing key acoustic segments is lacking in the speech domain. Furthermore, speech-enabled LLMs typically rely on encoders capable of processing only up to 30 seconds of audio (Liu et al., 2022; Chu et al., 2023; Radford et al., 2023; Liu et al., 2024), which limits their applicability for tasks requiring long-term context (Zhang et al., 2025c; Chen et al., 2025; Zhang et al., 2024b). This constraint poses a significant challenge for depression detection, as the task often necessitates analyzing extended acoustic patterns across entire conversations (Zhang et al., 2024d; Sun et al., 2022). Given these limitations, directly using speech LLMs is impractical for this domain, prompting the need to explore alternative approaches that leverage the advantages of text-based LLMs while effectively incorporating acoustic information from speech.

Building on the aforementioned limitations, a key research question emerges:

How can acoustic temporal patterns from speech be integrated into text-based LLMs for depression detection without training the LLMs?

To address this question, it is essential to develop a speech-based RAG framework. The central challenge lies in identifying the "region of interest" within acoustic patterns that can act as the optimal key for the RAG process. This region must encapsulate the distinguishing temporal and prosodic features that separate individuals with depression from healthy individuals, enabling the retrieval and generation components to better leverage acoustic information and achieve robust, scalable performance in depression detection. Prior research has established that acoustic landmarks provide critical features for depression detection systems, with particular emphasis on the discriminative power of their temporal patterns in differentiating between individuals with and without depression. (Zhang et al., 2024d; Huang et al., 2019a,c,b). Building on this insight, we designed the Speech Timing-based Retrieval-Augmented Generation framework, SpeechT-RAG, which leverages acoustic temporal patterns from landmarks to integrate speech information into text-based LLMs for depression detection without requiring additional training.

In developing our framework, we observed that these acoustic temporal patterns not only serve as effective features for depression detection but also naturally encode prediction reliability. While traditional methods for generating confidence scores often face challenges with LLMs (Yona et al., 2024a; Wu et al., 2024), the temporal patterns we identify through acoustic landmarks provide an interpretable basis for assessing prediction reliability. This insight enables us to develop an integrated

Table 1: Descriptions of the seven acoustic landmarks used in this study (Liu, 1996; Zhang et al., 2024d).

Landmark	Description
g	Onset (+) or offset (–) of vocal fold vibrations.
b	Onset (+) or offset (–) of turbulent noise during obstruent regions.
s	Onset (+) or offset (–) of nasal releases or closures.
v	Onset (+) or offset (–) of voiced frication.
p	Onset (+) or offset (–) of periodicity in voiced patterns.
f	Onset (+) or offset (–) of frication in unvoiced patterns.
j	Abrupt upward jump (+) or abrupt downward jump (–) in F0

approach where temporal information drives both detection and uncertainty quantification, enhancing system reliability in contexts where trustworthy decision-making is essential.

2 Preliminary

2.1 Acoustic Landmark

The concept of acoustic landmarks originates from studies on distinctive features (Garvin, 1953; Zhang et al., 2024c), emphasizing their role in phonetic contrasts and speech comprehension. Researchers have proposed that listeners rely on acoustic landmarks to extract essential cues for interpreting distinctive features, underscoring their significance in auditory processing (Liu, 1996). Over the years, acoustic landmarks have been extensively explored in various domains, including speech recognition (Liu, 1996; He et al., 2019), and more recently, in the field of depression detection (Huang et al., 2018, 2019a; Zhang et al., 2024d, 2025a), where they have shown potential for identifying speech patterns indicative of mental health conditions. Recent studies further suggest that the temporal information embedded within acoustic landmarks may hold particular value for applications in the health domain (Ishikawa et al., 2017; Huang et al., 2019a).

2.2 Computer Vision Retrieval-Augmented Generation

As illustrated in Figure 1, a common approach in computer vision for RAG involves extracting key entities or regions of interest from images (Jian et al., 2024; Xia et al., 2024), such as objects or salient areas, to serve as keys for retrieval and inputs for generation. This process effectively identifies distinctive features between images, enabling models to focus on task-relevant variations and improve interpretability. In contrast, speech lacks well-defined structures that can serve as analogous keys for RAG. Unlike visual data, where spatial features are clear, existing speech studies primarily rely on frame-level features or aggregated embeddings (Wu et al., 2023a, 2024), which often overlook nuanced temporal dynamics crucial for tasks like depression detection (Dineley et al., 2024). This gap highlights the need for innovative methods to define and extract meaningful keys from speech that align with the RAG paradigm.

3 SpeechT-RAG Framework

Our methodology involves three steps: First, we extract acoustic landmarks and construct sequential bigram pairs, embedding temporal information through their durations. Second, we identify statistically significant landmark pairs that exhibit notable differences between the two groups based on their duration statistics. Finally, we leverage these timing keys and text-based representations of temporal information to perform RAG with LLMs.

3.1 Landmark Extraction and Consecutive Bigram Construction

Previous studies have shown that temporal information is crucial for effective depression detection, as timing and rhythm often exhibit distinct patterns in affected individuals (Huang et al., 2019a,c; Dineley et al., 2024). To capture this essential temporal information, we leverage acoustic landmarks, which serve as indicators of key acoustic changes in speech, such as transitions between voiced and unvoiced regions, the onset of bursts, or sustained energy patterns in syllabic regions (Liu, 1996; Boyce et al., 2012; Zhang et al., 2024c). By utilizing the durations between consecutive acoustic landmarks, we extract the intrinsic temporal features embedded within speech signals, enabling a more precise analysis of timing patterns critical for identifying depression-specific characteristics.

Figure 2 stage one illustrates the process of extracting acoustic landmarks from speech, with a primary focus on the temporal intervals—*durations*—between consecutive landmarks as the core feature of interest. Table 1 lists the specific landmarks employed in this study, including **glottal (g)**, indicating vocal fold vibrations; **burst (b)**, representing plosive events; **syllabic (s)**, capturing vowels and sustained sonorants; **frication (f)** and **voiced frication (v)**, identifying unvoiced and voiced fricatives; and **periodicity (p)**, denoting recurring voiced patterns. **Jump (j)**, denotes abrupt change in F0. These landmarks collectively provide a comprehensive framework for analyzing the temporal dynamics of speech.

To encode the temporal relationships, we construct *landmark bigrams* by pairing consecutive landmarks within each utterance (Huang et al., 2019a,c). For two sequential landmarks l_i and l_{i+1}, the bigram is defined as $b_{i \to i+1} = (l_i, l_{i+1})$, and the corresponding duration is computed as:

$$d_{i \to i+1} = t(l_{i+1}) - t(l_i), \tag{1}$$

where $t(l_i)$ represents the timestamp of landmark l_i. This duration captures the temporal interval between two adjacent distinctive events, offering insights into speech timing. Unlike frame-based representations, this approach emphasizes the natural sequence and timing of events, characterizing the rhythmic properties of acoustic speech.

To analyze these durations, we calculated an enhanced set of statistical features for each landmark bigram. While previous work (Huang et al., 2019a) utilized basic statistics like mean and standard deviation, we introduce the interquartile range (IQR) to better capture the robustness of temporal variations, as IQR is less sensitive to outliers that commonly occur in natural speech patterns. Our statistical feature set includes the *mean*, *median*, *standard deviation*, *interquartile range (IQR)*, *minimum*, and *maximum* values:

$$\text{Statistical} = \{\mu, \text{med}, \sigma, \text{IQR}, \text{min}, \text{max}\}. \tag{2}$$

These features provide a detailed characterization of the temporal patterns in speech, preparing for distinguishing between healthy individuals and those with depression.

Table 2: Top 5 landmark pairs with the lowest p-values

Landmark Pair	U**-Statistic**	p**-Value**	**Health** μ	**Depression** μ
$+s - -v$	23042.5	0.00078	0.0795	0.0991
$+j - -v$	33793.5	0.00112	0.0309	0.0197
$-v - +j$	43035.5	0.00193	0.0725	0.0548
$-v - -j$	30640.5	0.01029	0.0523	0.0670
$+g - -v$	32691.0	0.01097	0.0100	0.0088

3.2 Identifying Distinctive Landmark Pairs

To identify the landmark pairs that exhibit statistically significant differences between healthy individuals and those with depression, We utilized the Mann-Whitney U test (McKnight and Najab, 2010) to systematically identify landmark bigram pairs whose duration distributions differ significantly between the non-depressed and depression groups. For a given bigram pair $b_{i \to i+1} = (l_i, l_{i+1})$, its aggregated duration set $D_{b_{i \to i+1}}$ consists of durations calculated across all samples. Specifically, we computed the set of statistical features $\{\text{mean}, \text{variance}, \text{interquartile range}, \ldots\}$ for $D_{b_{i \to i+1}}$. The Mann-Whitney U test was then applied to compare the feature distributions between the health and depression groups for each bigram pair:

$$U, p = \text{MannWhitneyU}(D^{\text{health}}_{b_{i \to i+1}}, D^{\text{depression}}_{b_{i \to i+1}}) \tag{3}$$

where p-values below 0.05 indicate significant differences in the temporal patterns of the bigram pair between the two groups.

Table 2 highlights the top five landmark pairs with the lowest p-values, indicating significant differences in their durations between the health and depression groups. These pairs, such as $+s - -v$ and $+j - -v$, demonstrate how temporal dynamics captured by landmark bigrams can effectively distinguish between the two groups. For subsequent tasks, we focus exclusively on those landmark bigrams with p-values less than 0.05, utilizing their statistical features to represent depression-specific characteristics.

3.3 SpeechT-RAG Framework Implementation

To effectively integrate speech timing information into the LLM for depression detection, we leverage a Speech Timing-Based Retrieval-Augmented Generation (RAG) framework as shown in Figure 2 stage 3. This framework selects representative examples from the training set based on mutual information (MI) scores and incorporates their temporal information into the text-based LLM.

MI measures the dependency between two random variables X and T. The MI for the random variable $X \in \mathbb{R}^{L\times D}$ and random variable $T \in \mathbb{R}^{L\times D}$ is expressed as:

$$\begin{aligned} I_i(X;T_i) &= H(X) - H(X \mid T_i) \\ &= D_{KL}\left(P(X,T_i) \parallel P(X) \otimes P(T_i)\right) \end{aligned} \tag{4}$$

where $H(X)$ is the entropy of X and $H(X \mid T)$ is the conditional entropy of X given T, D_{KL} denotes KL-divergence. Since directly estimating $I(X;T)$ is computationally intractable, MINE (Belghazi et al., 2018) approximates MI using a deep neural network, an approach that has found wide application across various domains (Ravanelli et al., 2020; Zhang et al., 2025b):

$$I_{\Theta}(X;T) = \mathbb{E}_{P(X,T)}[\psi_\theta] - \log(\mathbb{E}_{P(X)P(T)}[e^{\psi_\theta}]) \tag{5}$$

where ψ_θ is a statistics network parameterized by θ. Gradients of ψ_θ are computed by random batch sampling, ensuring efficient estimation of $I(X;T)$.

For each training sample $\mathbf{x}_{\text{train}}$, we compute its MI with a test sample $\mathbf{x}_{\text{test}}$, resulting in a set of MI scores. To ensure robustness, we calculate the average MI across test samples as:

$$\bar{I}(\mathbf{x}_{\text{train}}) = \frac{1}{M}\sum_{j=1}^{M} I(\mathbf{x}_{\text{test}}^{j}, \mathbf{x}_{\text{train}}) \tag{6}$$

where M is the number of test samples. Based on these MI scores, the top n training examples most similar to the test sample are selected, including n examples each from the health and depression classes:

$$\mathcal{R} = \mathcal{H} \cup \mathcal{D}, \quad |\mathcal{H}| = |\mathcal{D}| = n, \tag{7}$$

where $\mathcal{H}$ and $\mathcal{D}$ denote the selected health and depression examples, respectively.

To incorporate timing information into the LLM, the temporal features (e.g., mean, variance) of landmark bigram durations are converted into a structured text format. Each timing sequence is formatted as:

$$\mathbf{t}_{\text{sample}} = \text{Format}(\mathbf{d}_{\text{bigram}}) \tag{8}$$

where $\mathbf{d}_{\text{bigram}}$ represents the statistical features of bigram durations. For example:

$$\mathbf{t}_{\text{sample}} : \text{+s–v (mean: 0.08, var: 0.01)}, \tag{9}$$

Here, *+s–v* represents a transition from a syllabic onset to a voiced frication.

The formatted representations of the retrieved examples are concatenated with the timing information of the test sample to construct a prompt:

$$\mathbf{P} = \mathbf{t}_{\text{example}_1} + \mathbf{t}_{\text{example}_2} + \cdots + \mathbf{t}_{\text{test}} \tag{10}$$

where $+$ denotes concatenation. The prompt, along with an instruction, is fed into the LLM for classification. The LLM generates predictions for the test sample as either *Health* or *Depression*.

4 Confidence Score Estimation

Confidence estimation is a critical component of health-related AI systems, offering a measure of reliability for predictions, which is essential for clinical decision-making (Edin et al., 2024; Kang et al., 2024). While traditional machine learning systems frequently employ confidence scoring techniques (Wu et al., 2024), their application to LLMs remains challenging (Yona et al., 2024b; Chaudhry et al., 2024). To address this, we propose a framework based on Gaussian Process Classifiers (GPCs) that leverages speech timing information, specifically the temporal features of landmark bigrams with p-values below 0.05, to predict confidence scores.

Gaussian Process Classifier Using Speech Timing Information. The GPC models the relationship between input features and class labels within a probabilistic framework. Here, the input feature vector $\mathbf{x} \in \mathbb{R}^d$ is constructed using the statistical characteristics (e.g., mean, variance) of durations from landmark bigrams identified as significant ($p < 0.05$) in the earlier analysis, where d is the dimensionality of the feature vector. The class label $y \in \{0, 1\}$ denotes either the depression or health category.

The GPC predicts class probabilities $P(y \mid \mathbf{x})$ using a radial basis function (RBF) kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right) \tag{11}$$

where σ^2 is the signal variance and l is the length scale, which controls the smoothness of the kernel function. This kernel facilitates capturing the nonlinear relationships in the timing features, allowing robust probabilistic modeling of the speech data.

For a test feature vector $\mathbf{x}_{\text{test}}$, the confidence score C is derived from the predicted probabilities as:

$$C = \max(P(y = 0 \mid \mathbf{x}_{\text{test}}), P(y = 1 \mid \mathbf{x}_{\text{test}})) \tag{12}$$

Expected Calibration Error (ECE) ECE is a widely used metric to evaluate the quality of confidence scores by measuring their alignment with observed accuracies (Chaudhry et al., 2024; Wu et al., 2024). For n test samples, ECE is computed as:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} |\text{conf}(B_b) - \text{acc}(B_b)| \quad (13)$$

where B_b denotes the set of samples in the b-th bin, and $|B_b|$ is the number of samples in the bin. A smaller ECE value indicates better model calibration.

5 Experiments

5.1 Experiments Setup

Dataset The DAIC-WOZ dataset (DeVault et al., 2014) is widely regarded as a benchmark resource for depression detection tasks. It comprises 189 recordings of clinical interviews conducted between interviewers and patients. Within the training set, 30 out of 107 interviews are labeled as depressed, while the development set contains 35 interviews, with 12 categorized as depressed. Following the practices of prior research, e.g. (Gong and Poellabauer, 2017; Shen et al., 2022; Wu et al., 2022, 2023a), our experimental results are evaluated on the development set.

Model Configurations We selected our evaluation models based on parameter efficiency and context processing capabilities. From the Llama2 family (Touvron et al., 2023), we included both base models (Llama2-7B, Llama2-13B) and their instruction-tuned variants (Llama2-7B Chat, Llama2-13B Chat) to assess the impact of instruction tuning. We also incorporated Llama3 models (Dubey et al., 2024) (Llama3-8B-Instruct and Llama3-8B) for their enhanced context processing abilities. For Text-RAG baseline, we employed the SentenceTransformer (Reimers and Gurevych, 2019) model `all-MiniLM-L6-v2` to compute embeddings and rank examples using cosine similarity, retrieving the top n examples (n = 1 or 2) from the training set.

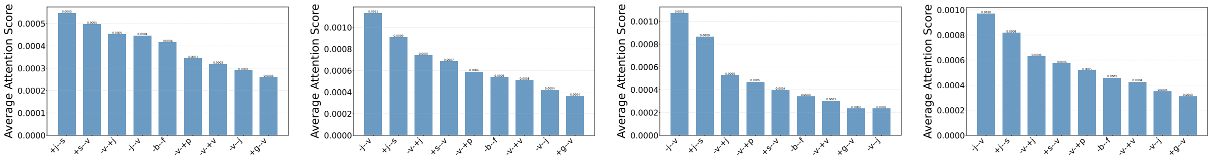
5.2 Main Results

Table 3 demonstrates the superior performance of our SpeechT-RAG approach over traditional Text-RAG methods in depression detection. As shown by the F1 scores, SpeechT-RAG exhibits consistent improvement with an increasing number of retrieved examples, highlighting the importance of temporal speech patterns in distinguishing between depressed and healthy individuals. In contrast, Text-RAG's performance deteriorates with additional retrieved examples, suggesting that text-based retrieval introduces noise that compromises classification accuracy.

Table 3: F1 scores across different large language models and retrieval methods. The results for Speech Self-Supervised (SSL) models and Llama2 Fine-tune are taken from (Wu et al., 2023a) and (Zhang et al., 2024d), which applied extensive data augmentation specifically for depression detection. The "Method" column specifies the retrieval or input type, while the "Examples" column indicates the number of retrieved examples used.

Model	Method	Examples	F1-avg	F1-max	F1-std
Speech SSL Baselines	Wav2Vec 2.0	-	0.627	0.667	0.043
	HuBERT	-	0.667	0.762	0.052
	WavLM	-	0.700	0.750	0.024
Llama2 7B Chat	Zero-shot	0	0.195	0.207	0.023
	Fine-tune	-	0.488	-	-
	Timing-RAG	2	0.563	0.600	0.032
Llama2 7B	Zero-shot	0	0.173	0.182	0.011
	Fine-tune	-	0.578	-	-
	Timing-RAG	2	0.548	0.571	0.030
Llama2 13B Chat	Zero-shot	0	0.186	0.191	0.005
	Fine-tune	-	0.545	-	-
	Timing-RAG	2	0.528	0.533	0.007
Llama2 13B	Zero-shot	0	0.249	0.300	0.037
	Fine-tune	-	0.636	-	-
	Timing-RAG	2	0.516	0.581	0.058
Llama3 8B	Zero-shot	0	0.528	0.537	0.008
	Text-RAG	1	0.458	0.487	0.020
	Text-RAG	2	0.292	0.316	0.038
	Timing-RAG	1	0.507	0.571	0.047
	Timing-RAG	2	0.601	0.640	0.026
	Timing-RAG	4	0.651	0.692	0.048
Llama3 8B Instruct	Zero-shot	0	0.517	0.537	0.021
	Text-RAG	1	0.627	0.643	0.021
	Text-RAG	2	0.304	0.316	0.008
	Timing-RAG	1	0.497	0.500	0.007
	Timing-RAG	2	0.624	0.625	0.002
	Timing-RAG	4	0.692	0.733	0.049

The effectiveness of SpeechT-RAG is further validated across different model architectures. Despite Llama2's context window limitation allowing only two retrieved examples, it achieves performance comparable to fine-tuned models — establishing SpeechT-RAG as a resource-efficient alternative for scenarios with limited data or computational constraints. Leveraging Llama3's expanded context capacity, SpeechT-RAG demonstrates even stronger performance gains by effectively incorporating up to four examples, underscoring the scalability of our approach.



(a) Early Layers (2-4)

(b) Middle Layers

(c) Final 4 Layers

(d) All Analyzed Layers

Figure 3: Attention scores across different landmark pairs for varying Transformer layers. Subfigures represent attention scores computed for early, middle, final, and all layers.

Table 4: Comparison of ECE results for Llama3 models and baselines. The results for MCDP (Gal and Ghahramani, 2016) and PWLM (Wu et al., 2024) baselines are derived from (Gal and Ghahramani, 2016) and (Wu et al., 2023a), respectively. The ECE values are calculated based on overall accuracy for the depression detection task. Mean ECE and standard deviation (Std) are reported.

Model	**Mean ECE (↓)**	**ECE Std**
PWLM	0.183	0.009
Llama3 8B MCDP	0.349	0.002
Llama3 8B Instruct MCDP	0.340	0.003
Llama3 8B	0.0674	0.0481
Llama3 8B Instruct	0.0276	0.0274

Table 4 presents the Expected Calibration Error (ECE) analysis, comparing our Timing-RAG approach with two established calibration methods. The first baseline, PWLM (Wu et al., 2023a), is a self-supervised speech model specifically trained for depression detection, with results cited from the original work. The second baseline, MCDP (Wu et al., 2024),employs Monte Carlo Dropout during prediction to estimate model uncertainty and was evaluated using both Llama3 8B and Llama3 8B Instruct models.

Our experimental results demonstrate the superior calibration performance of Timing-RAG. While MCDP achieves reasonable calibration with ECE values of 0.33 and 0.34, our Timing-RAG implementation with Llama3 8B and Llama3 8B Instruct models achieves substantially lower mean ECE values of 0.0674 and 0.0276, respectively. The consistently low standard deviations across trials further validate the reliability of our approach. These results underscore how incorporating speech timing information through Timing-RAG significantly enhances both the calibration quality and prediction reliability in LLM-based depression detection.

6 Discussion

6.1 Landmark Pair Importance Analysis

To gain a deeper understanding of how the Llama3 Instruct model leverages speech timing information for decision-making, we analyzed its attention mechanisms across different stages of processing. We selected the Llama3 Instruct model for this analysis based on its demonstrated stability and consistently high F1 scores across multiple experiments. The model's layers were categorized into three distinct groups: early layers (layers 2-4), middle layers (20%-80% of total layers), and final layers (the last 4 layers), enabling us to track the evolution of attention patterns throughout the model's processing pipeline.

For each layer group, we computed attention scores by identifying and aggregating attention weights assigned to tokens corresponding to landmark pairs in the input sequence. To quantify the importance of each landmark pair, we developed a scoring mechanism that considers both the magnitude and consistency of attention:

$$\text{Score} = \mu \times (1 + 0.5 \cdot \sigma) \tag{14}$$

where μ represents the mean attention score and σ denotes the standard deviation across layers. This formulation extends beyond simple averaging by incorporating variance information, allowing us to distinguish between two scenarios: landmark pairs that maintain steady, high attention across layers (indicating consistent relevance to the model's decision process) versus those that receive sporadic high attention but may not be consistently meaningful. The scaling factor of 0.5 was empirically chosen to balance the influence of consistency versus absolute attention magnitude.

Building upon the earlier analysis of the model's attention mechanism, we observed that in the early layers, the Llama3 Instruct model prioritizes landmark pairs such as $+s \rightarrow -v$ and $-v \rightarrow +j$, which closely align with the most significant land-

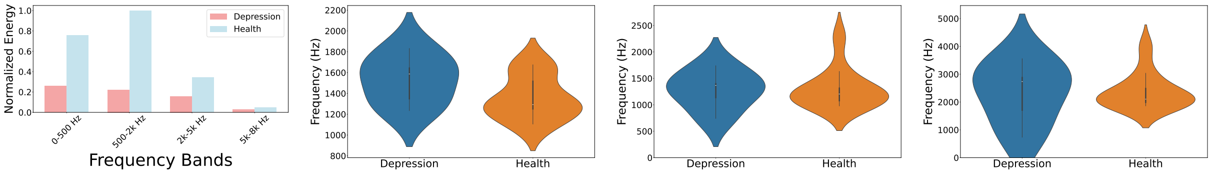


Figure 4: Acoustic analysis of speech segments for landmark pair (+s, -v): (a) Energy distribution across frequency bands; (b) Spectral bandwidth distribution; (c) Spectral centroid distribution; and (d) Spectral rolloff distribution, comparing depression and health groups.

mark pairs identified in our speech signal analysis, based on the lowest p-values. This similarity suggests that, similar to speech self-supervised learning (SSL) models (Hsu et al., 2021; Pasad et al., 2021; Chung et al., 2021), the Llama3 Instruct model may focus on low-level features in its early layers. This pattern suggests that both speech pre-trained models and text pre-trained models may first focus on low-level information before shifting to high-level information in their decision-making process for depression detection. In the middle and final layers, the Llama3 Instruct model focuses on similar landmark pairs, indicating a consistent attention pattern as the model processes speech features. These landmark pairs, such as $+s \to -v$ and $-v \to +j$,often associated with more abrupt changes in speech. The focus on these landmark pairs suggests that, in the deeper layers, the model emphasizes landmarks that represent more dynamic and notable speech features.

6.2 Acoustic Validation of Detection-Critical Landmark Pairs

Based on our analyses of both LLM attention patterns and statistical tests, we observed a consistent emphasis on the landmark pair (+s, -v), which appears to be particularly informative for depression detection. This finding motivates a detailed acoustic investigation of speech segments marked by this landmark pair. As shown in Figure 4, we analyzed four key spectral features. The energy band distribution analysis reveals that healthy subjects maintain notably higher energy levels across all frequency bands, with the most pronounced difference in the speech fundamental frequency range (500Hz-2kHz). This difference suggests reduced vocal energy in depressed speech, particularly in frequencies crucial for speech articulation. For each band b, we calculate the normalized energy E_b as:

$$E_b = \frac{1}{N} \sum_{f \in b} |X(f)|^2 \tag{15}$$

The spectral bandwidth distribution shows that depressed subjects exhibit a more compressed frequency spread, indicating less variability in their vocal frequency components. This reduced bandwidth suggests a more monotonic speech pattern, which aligns with clinical observations of reduced prosodic variation in depression (Peper and Lin, 2012; Quatieri and Malyska, 2012). The spectral centroid distribution further supports this finding, showing that depressed speech tends to have lower centroid values. This indicates that the center of mass of the frequency spectrum is shifted towards lower frequencies, suggesting a less "bright" or more dampened vocal quality characteristic of depressed speech. The spectral roll-off analysis, representing the frequency below which 85% of the spectral energy is contained, reveals that depressed speech consistently shows lower roll-off frequencies. This indicates a concentration of energy in lower frequency bands, further supporting the observation of reduced vocal expressiveness and energy in depressed speech patterns.

Collectively, these acoustic findings suggest that our landmark-based timing system captures specific energy patterns at critical acoustic transitions. The differences in energy distribution and frequency composition during the (+s, -v) segments demonstrate that these temporal transition points serve as meaningful acoustic indicators for depression detection.

7 Conclusion

In this paper, we have introduced SpeechT-RAG, a novel Retrieval-Augmented Generation framework that leverages acoustic temporal patterns for depression detection and uncertainty assessment. Through systematic analysis of LLM attention mechanisms and speech characteristics, we demonstrated how temporal information embedded in acoustic landmark pairs can simultaneously serve two critical functions: capturing depression-related patterns (such as reduced energy levels and diminished vocal expressiveness) and providing a natural basis for assessing prediction reliability. The dual utility of these temporal features enables our unified framework to not only achieve strong detection performance but also offer interpretable confidence estimation without additional computational overhead. These findings demonstrate how domain-specific temporal patterns can enhance both the accuracy and reliability of LLM frameworks, advancing the development of trustworthy systems.

Limitations

A key limitation of our study is its reliance on the DAIC-WOZ dataset. While this dataset represents the current standard in multimodal depression recognition research and remains the only publicly accessible resource for speech-based depression analysis, this singular focus may impact the generalizability of our findings. The scarcity of available datasets in this domain stems from the significant privacy and ethical considerations surrounding mental health data collection. Nevertheless, our approach of conducting comprehensive analysis on this widely-used benchmark aligns with established research practices in the field of speech depression detection.

Ethics Statement

Our research utilizes the DAIC-WOZ dataset, which has undergone rigorous de-identification procedures to ensure participant privacy protection. However, we acknowledge important ethical considerations regarding the deployment of our system. While our approach demonstrates improved accuracy in processing speech patterns for depression detection, the current performance level may not yet meet the stringent requirements for clinical applications. Like all machine learning models, our system may exhibit inherent biases that could lead to incorrect observations or classifications. Therefore, we emphasize that any practical implementation of this technology should be conducted under careful professional supervision, with our system serving as a supportive tool rather than a primary diagnostic mechanism.

Acknowledgement

This work was supported by Australian Research Council Discovery Project DP230101184.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.

Suzanne Boyce, Harriet Fell, and Joel MacAuslan. 2012. Speechmark: Landmark detection tool for speech analysis. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. 2024. Finetuning language models to emit linguistic expressions of uncertainty. *arXiv preprint arXiv:2409.12180*.

Moran Chen, Qiquan Zhang, Mingjiang Wang, Xiangyu Zhang, Hexin Liu, Eliathamby Ambikairaiah, and Deying Chen. 2025. Selective state space model for monaural speech enhancement. *IEEE Transactions on Consumer Electronics*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Yu-An Chung, Yonatan Belinkov, and James Glass. 2021. Similarity analysis of self-supervised speech representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3040–3044. IEEE.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Judith Dineley, Ewan Carr, Lauren L White, Catriona Lucas, Zahia Rahman, Tian Pan, Faith Matcham, Johnny Downs, Richard J Dobson, Thomas F Quatieri, et al. 2024. Variability of speech timing features across repeated recordings: a comparison of open-source extraction techniques. In *Interspeech 2024*, pages 2015–2019. ISCA.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Miami, Florida, USA. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Paul L Garvin. 1953. Preliminaries to speech analysis: The distinctive features and their correlates.

Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, pages 69–76.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Di He, Xuesong Yang, Boon Pang Lim, Yi Liang, Mark Hasegawa-Johnson, and Deming Chen. 2019. When ctc training meets acoustic landmarks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5996–6000. IEEE.

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

Zhaocheng Huang, Julien Epps, and Dale Joachim. 2019a. Investigation of speech landmark patterns for depression detection. *IEEE transactions on affective computing*, 13(2):666–679.

Zhaocheng Huang, Julien Epps, and Dale Joachim. 2019b. Speech landmark bigrams for depression detection from naturalistic smartphone speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5856–5860. IEEE.

Zhaocheng Huang, Julien Epps, Dale Joachim, and Michael Chen. 2018. Depression detection from short utterances via diverse smartphones in natural environmental conditions. In *INTERSPEECH*, pages 3393–3397.

Zhaocheng Huang, Julien Epps, Dale Joachim, and Vidhyasaharan Sethu. 2019c. Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection. *IEEE Journal of selected topics in Signal Processing*, 14(2):435–448.

Keiko Ishikawa, Joel MacAuslan, and Suzanne Boyce. 2017. Toward clinical application of landmark-based speech analysis: Landmark expression in normal adult speech. *The Journal of the Acoustical Society of America*, 142(5):EL441–EL447.

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956.

Migyeong Kang, Goun Choi, Hyolim Jeon, Ji Hyun An, Daejin Choi, and Jinyoung Han. 2024. CURE: Context- and uncertainty-aware mental disorder detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17924–17940, Miami, Florida, USA. Association for Computational Linguistics.

Adi Lahat, Eyal Shachar, Benjamin Avidan, Zina Shatz, Benjamin S Glicksberg, and Eyal Klang. 2023. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Scientific reports*, 13(1):4164.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hexin Liu, Leibny Paola Garcia, Andy W. H. Khong, Eng Siong Chng, Suzy J. Styles, and Sanjeev Khudanpur. 2022. Efficient self-supervised learning representations for spoken language identification. *IEEE J. Sel. Topics Signal Process.*, 16(6):1296–1307.

Hexin Liu, Xiangyu Zhang, Haoyang Zhang, Leibny Paola Garcia, Andy W. H. Khong, Eng Siong Chng, and Shinji Watanabe. 2024. Aligning speech to languages to enhance code-switching speech recognition. *arXiv preprint arXiv:2403.05887*.

Sharlene A Liu. 1996. Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America*, 100(5):3417–3430.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

Namkee Oh, Gyu-Seong Choi, and Woo Yong Lee. 2023. Chatgpt goes to the operating room: evaluating gpt-4 performance and its potential in surgical education and training in the era of large language models. *Annals of Surgical Treatment and Research*, 104(5):269.

Julia Ohse, Bakir Hadžić, Parvez Mohammed, Nicolina Peperkorn, Michael Danner, Akihiro Yorita, Naoyuki Kubota, Matthias Rätsch, and Youssef Shiban. 2024. Zero-shot strike: Testing the generalisation capabilities of out-of-the-box llm models for depression detection. *Computer Speech & Language*, 88:101663.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Erik Peper and I-Mei Lin. 2012. Increase or decrease depression: How body postures influence your energy level. *Biofeedback*, 40(3):125–130.

Thomas F Quatieri and Nicolas Malyska. 2012. Vocal-source biomarkers for depression: A link to psychomotor activity. In *Interspeech*, volume 2, pages 1059–1062.

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

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

Hao Sun, Yen-Wei Chen, and Lanfen Lin. 2022. Tensorformer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection. *IEEE Transactions on Affective Computing*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta. 2013. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48.

Wen Wu, Mengyue Wu, and Kai Yu. 2022. Climate and weather: Inspecting depression detection via emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6262–6266. IEEE.

Wen Wu, Chao Zhang, and Philip C Woodland. 2023a. Self-supervised representations in speech-based depression detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wen Wu, Chao Zhang, and Philip C Woodland. 2024. Confidence estimation for automatic detection of depression and alzheimer's disease based on clinical interviews. In *Proc. Interspeech 2024*, pages 3160–3164.

Yuhan Wu, Yuanyuan Xu, Xuemin Lin, and Wenjie Zhang. 2023b. A holistic approach for answering logical queries on knowledge graphs. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 2345–2357. IEEE.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093.

Gal Yona, Roee Aharoni, and Mor Geva. 2024a. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024b. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. 2024a. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Xiangyu Zhang, Beena Ahmed, and Julien Epps. 2025a. Why pre-trained models fail: Feature entanglement in multi-modal depression detection. *arXiv preprint arXiv:2503.06620*.

Xiangyu Zhang, Daijiao Liu, Hexin Liu, Qiquan Zhang, Hanyu Meng, Leibny Paola Garcia Perera, EngSiong Chng, and Lina Yao. 2024b. Speaking in wavelet domain: A simple and efficient approach to speed up speech diffusion model. In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing, pages 159–171.

Xiangyu Zhang, Daijiao Liu, Tianyi Xiao, Cihan Xiao, Tuende Szalay, Mostafa Shahin, Beena Ahmed, and Julien Epps. 2024c. Auto-landmark: Acoustic landmark dataset and open-source toolkit for landmark extraction. *arXiv preprint arXiv:2409.07969*.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024d. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 146–158.

Xiangyu Zhang, Jianbo Ma, Mostafa Shahin, Beena Ahmed, and Julien Epps. 2025b. Rethinking mamba in speech processing by self-supervised models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. 2025c. Mamba in speech: Towards an alternative to self-attention. *IEEE Transactions on Audio, Speech and Language Processing*.

Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Haishuai Wang, and Björn Schuller. 2020. Hierarchical attention transfer networks for depression assessment from speech. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7159–7163. IEEE.

A Example Prompts

A.1 Zero Shot Prompt Example

```
You are a mental health expert.

Your task is to classify if a patient is
     depressed or healthy based on their
     dialogue.

You must respond with ONLY ONE WORD:
    either 'Depressed' or 'Health'.

Conversation:
{dialogue}

Diagnosis:
```

A.2 Text Rag Prompt Example

```
You are a mental health expert.

Your task is to classify if a patient is
     depressed or healthy based on their
     dialogue.

You must respond with ONLY ONE WORD:
    either 'Depressed' or 'Health'.

"Case {idx}:
{example['dialogue']

Classification: {label}

Here are some example cases with their
    classifications:
{dialogue}

Now classify the following case with
    ONLY ONE WORD (Depressed or Health):

{dialogue}

Classification:
```

A.3 Speech-Timing Rag Prompt Example

```
"The task is to classify patients as '
    Depression' or 'Health' based on
    their statistical feature patterns.
    "

"Each example below shows a sequence of
    statistical values followed by the
    correct classification Class."

format_example(item['
    bigram_durations_statics'], item['
    label'])

Class:
```