When LLMs Meet Acoustic Landmarks: An Efficient Approach to Integrate Speech into Large Language Models for Depression Detection

Xiangyu Zhang[1], **Hexin Liu**[2], **Kaishuai Xu**[3],
Qiquan Zhang[1], **Daijiao Liu**[1], **Beena Ahmed**[1], **Julien Epps**[1]
The University of New South Wales[1]
Nanyang Technological University[2] The Hong Kong Polytechnic University[3]

Abstract

Depression is a critical concern in global mental health, prompting extensive research into AI-based detection methods. Among various AI technologies, Large Language Models (LLMs) stand out for their versatility in mental healthcare applications. However, their primary limitation arises from their exclusive dependence on textual input, which constrains their overall capabilities. Furthermore, the utilization of LLMs in identifying and analyzing depressive states is still relatively untapped. In this paper, we present an innovative approach to integrating acoustic speech information into the LLMs framework for multimodal depression detection. We investigate an efficient method for depression detection by integrating speech signals into LLMs utilizing Acoustic Landmarks. By incorporating acoustic landmarks, which are specific to the pronunciation of spoken words, our method adds critical dimensions to text transcripts. This integration also provides insights into the unique speech patterns of individuals, revealing the potential mental states of individuals. Evaluations of the proposed approach on the DAIC-WOZ dataset reveal state-of-the-art results when compared with existing Audio-Text baselines. In addition, this approach is not only valuable for the detection of depression but also represents a new perspective in enhancing the ability of LLMs to comprehend and process speech signals.

1 Introduction

Depression, a common mental disorder affecting 10-15% of the global population, is characterized by persistent low mood, loss of interest, and lack of energy, making it a prevalent and costly illness (Walker et al., 2018). Given the time-consuming, expensive, and sometimes ineffective nature of traditional depression treatment methods, a growing number of researchers are turning their attention to developing automated depression detection systems. Concurrently, Large language models (LLMs) have recently demonstrated remarkable success across a variety of tasks (Chowdhery et al., 2023; Touvron et al., 2023). These large language models have been applied to various healthcare issues, including general surgery (Oh et al., 2023), dementia diagnosis (Wang et al., 2023), and gastroenterology (Lahat et al., 2023) and achieved excellent results. However, their main limitation stems from their sole reliance on textual input, which limits their full potential. Simultaneously, the use of Large Language Models (LLMs) in depression detection remains largely unexplored. In particular, there has been no effort to integrate speech—despite growing evidence that speech signals can reveal indicators of depression (Wu et al., 2023; Huang et al., 2019a)—into these LLMs, an advancement that could greatly improve their effectiveness in identifying depression (Zheng et al., 2023).

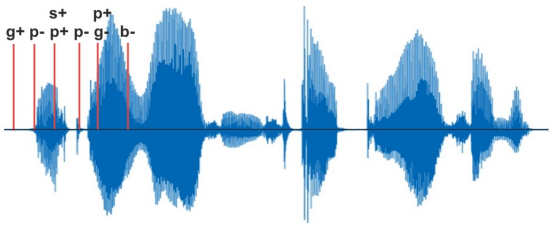


Figure 1: Example of Acoustic Landmark (2-gram concat landmark (g+p-), (s+p+), (p+,p-), ..., (g-b-)), Landmarks are extracted from abrupt changes in the speech signal. They can discretize speech into a series of tokens that possess linguistic significance.

One of the key approaches to incorporating speech signals into LLMs is through the discretization of speech. However, the current landscape of speech discretization, heavily reliant on deep learning techniques (Zeghidour et al., 2021; Défossez et al., 2022), faces significant challenges due to its considerable GPU memory requirements. This is particularly problematic in the field of depres-

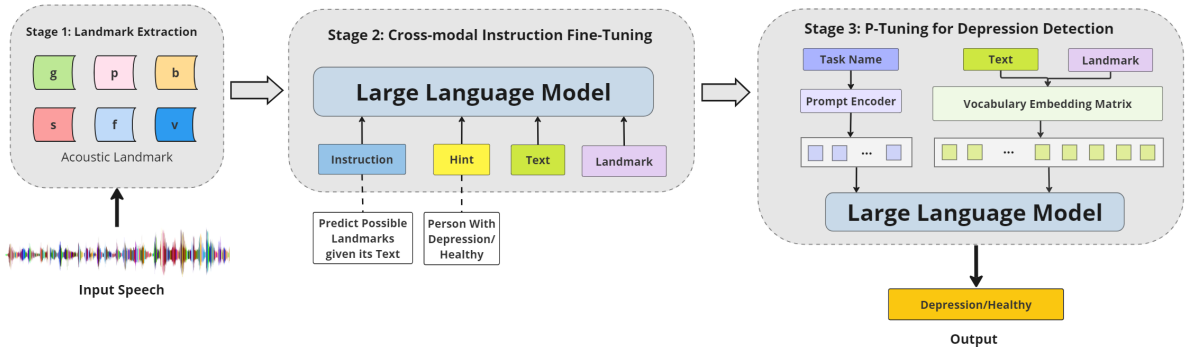


Figure 2: Overview of LLM-Landmark Depression Detection Pipeline, broadly categorized into three stages: landmark detection (on the left), cross-modal instruction fine-tuning (in the middle), and P-tuning for depression detection (on the right).

sion detection, where data often consists of lengthy conversations (DeVault et al., 2014). The need for completed conversations is vital for accurate depression detection (Wu et al., 2023; Sun et al., 2022), rendering the existing deep learning-based methods impractical for such applications. For this purpose, it is necessary to find an efficient approach that allows for the discretization of speech with reduced GPU memory usage.

Acoustic landmarks represent event markers intricately linked with the articulation of speech, forming a concise alternative framework for speech processing (Liu, 1996; Stevens, 2002). This approach emphasizes the analysis of abrupt acoustic changes at the subsegmental level, thereby providing a succinct and precise phonetic description of language. These landmarks, characterized by their binary values, establish a minimal yet effective set for differentiating each language segment from others. They maintain a direct and significant relationship with acoustic properties and articulation (including individual pronunciation), ensuring discernibility despite unwanted variability introduced by diverse hardware and environmental backgrounds (Huang et al., 2018, 2019b). Their discrete nature not only allows for efficient integration into large language models but also offers a viable alternative for understanding speech signals in depression detection, bypassing the limitations of current deep learning-based techniques. This innovative approach promises a more feasible and resource-efficient pathway for analyzing complex speech patterns in mental health diagnostics.

In this paper, we introduce a novel multimodal approach to depression detection, utilizing a combination of acoustic landmarks and large language models. We investigate the properties of large language models at various stages and under different conditions after integrating landmark-based speech information. We investigate how LLMs learn speech landmarks and assess the impact of conversational fine-tuning on the performance of LLMs in tasks related to depression detection.

In summary, our contributions include the following:

- To the best of our knowledge, this is the first study to apply LLMs to **multimodal** depression detection and the inaugural effort to integrate speech information into LLMs for this purpose. We proposed a new baseline for the application of LLMs in the field of automatic depression detection.
- Compared with prior baseline audio-text methods (Wu et al., 2023), our approach not only achieved SOTA performance but also involved a comprehensive analysis of the properties of LLMs post the integration of landmarks.
- Unlike previous deep learning-based methods for aiding LLMs in understanding speech, we explored a new, more efficient approach to enable LLMs to process speech signals. This novel method opens up a potentially groundbreaking direction for enhancing LLMs' comprehension of speech.

2 Related Work

2.1 Large Language Models

Large language models have achieved success in natural language processing and have been extended to encompass computer vision and speech signal processing (Brown et al., 2020; Touvron et al., 2023; Li et al., 2023b; Liu et al., 2024). However, there is a significant gap in research aimed at enabling LLMs to comprehend speech efficiently.

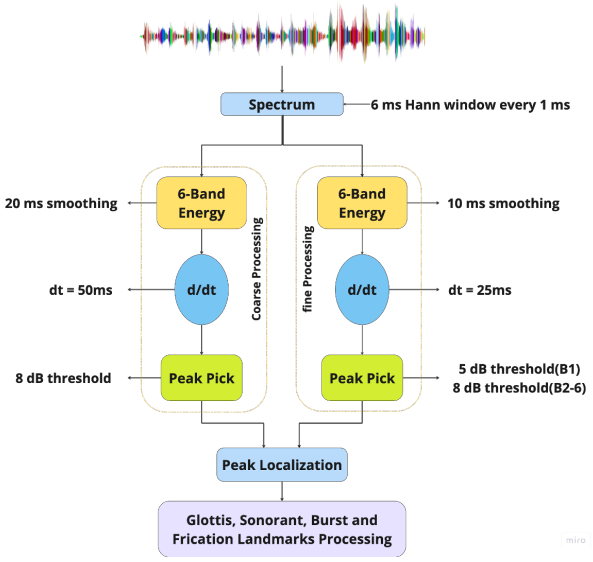


Figure 3: Landmark Detection Filter

Parameter-efficient fine-tuning refers to selectively updating a small subset of the model's parameters or adding lightweight trainable layers, to customize the model for specific tasks or domains with reduced computational overhead. Existing works employed low-rank adaptation (LoRA) to fine-tune LLM efficiently. LoRA reduces computational complexity by freezing the pre-trained LLM and injecting trainable rank decomposition matrices A and B into its transformer-based layers (Hu et al., 2022). The forward pass is subsequently defined as the linear combination of those from the pre-trained model and from the trained decomposed matrices A and B.

2.2 Acoustic Landmarks

The concept of acoustic landmarks originally stems from research on distinctive features (Garvin, 1953; Zhang et al., 2024a). Some researchers posit that for certain phonetic contrasts, a listener relies on acoustic landmarks to gather the necessary acoustic cues for deciphering the underlying distinctive features (Liu, 1996). This perspective highlights the importance of these landmarks in the auditory processing and interpretation of speech. Subsequent research has utilized acoustic landmarks for applications in speech recognition (Liu, 1996; He et al., 2019) as well as in addressing mental health-related problems (Huang et al., 2018, 2019a). Although different scholars have slightly varied definitions of acoustic landmarks, Joel and colleagues (Boyce et al., 2012) expanded upon Liu's paper (Liu, 1996) by releasing a MATLAB version of a landmark detection toolkit, which has become the most widely used version of landmark technology.

2.3 Automatic Depression Detection

The use of AI technology for depression detection has been developing for many years. Some researchers (Cummins et al., 2011; Huang et al., 2018, 2019a) have utilized traditional methods such as Support Vector Machines (SVMs) (Noble, 2006) for depression detection. With the advancement of deep learning technologies (Gulati et al., 2020; Zhang et al., 2024c), an increasing number of researchers have been experimenting with deep learning approaches for depression detection. Zhao and others have explored the use of transformer models for processing speech inputs in depression detection (Zhao et al., 2020). Shen and colleagues have employed BI-LSTM architectures, combining text and speech for this purpose (Shen et al., 2022). Further extending these techniques, Wu (Wu et al., 2023) utilized speech self-supervised models (Chen et al., 2022; Hsu et al., 2021; Liu et al., 2022) and integrated them with RoBERTa (Liu et al., 2019) for a more comprehensive text-audio multimodal approach to depression detection.

3 Methodology

3.1 Overview

Our methodology, detailed in Figure 2, encompasses a three-step training process. The first phase involves extracting acoustic landmarks from speech and conducting an array of data processing operations. Subsequently, in the Cross-modal Instruction Fine-Tuning phase, we engage the LLM in learning the nuances and characteristics of acoustic landmarks. The culminating phase is the P-Tuning process, wherein the LLM is meticulously trained to apply its understanding to diagnose depression.

3.2 Landmarks Extraction and Data Preprocessing

3.2.1 Landmarks Extraction

Figure 1 illustrates an example of acoustic landmarks, where speech signals are discretized into a series of symbols that carry linguistic relevance. Table 1 details the specific acoustic landmarks utilized in our study. Diverging from Liu's paper (Liu, 1996), our research also pays attention to frication, voice frication, and periodicity.

Our method primarily draws inspiration from Joel's (Boyce et al., 2012) and Liu's (Liu, 1996) work. However, since they have not open-sourced

Landmark	Description
g	vibration of vocal folds start (+) or end (–)
b	onset (+) or offset (–) of existence of turbulent noise during obstruent regions
s	releases (+) or closures (–) of a nasal
v	voiced frication onset (+) or offset (–)
p	periodicity start (+) or end (–)
f	frication onset (+) or offset (–)

Table 1: Description of the six landmarks investigated.

their code, many of their approach's details remain unknown. In the following section, We introduce our Python-based landmark detection algorithm, developed to address these gaps and to adapt the conceptual framework to our specific requirements. Initially, the spectrogram is divided into six frequency bands. Landmarks are identified through energy changes within these six bands, using a two-pass strategy. Different landmarks are determined by either a single band or a combination of multiple bands (Liu, 1996). This approach is visually represented by the two parallel branches emanating from the spectrogram block in Figure 3.

The detection algorithm for **Glottal (g)**, **Burst (b)**, and **Syllabic (s)** landmarks is fundamentally aligned with Liu's approach (Liu, 1996). However, diverging from Liu's method, we employ 5dB and 8dB as threshold values because of different smoothing methods between Python and Matlab. Additionally, considering that the opening and closing of the glottis occur in pairs, We implemented dynamic programming to ensure that g landmarks appear in pairs, thus enhancing the physiological accuracy of our detection.

Our methodology for identifying **f+** and **v+** landmarks involves detecting a 6 dB power increase in at least three high-frequency bands (bands 4-6), and a power decrease in low-frequency bands (bands 2 and 3). For **f-** and **v-**, the criteria are reversed: a 6 dB power decrease in the same high-frequency bands and a power increase in the low-frequency bands. The distinguishing factor here is that frication landmarks are detected within unvoiced segments (b landmark), while voiced frication landmarks are sought in voiced segments (s landmark).

Regarding the detection of the **periodicity (p)** landmarks, we perform autocorrelation calculations on the audio frame to identify repetitive or periodic patterns in the data. For a detailed description of our landmark detection algorithm, please refer to Appendix A.

3.2.2 Data Augmentation and Processing

Depression assessments are commonly conducted through clinical interviews, with each session receiving a singular label. This labeling method, when applied to a given dataset size, leads to fewer samples in datasets compared with the much larger number of utterances and frames typically encountered in other speech-related tasks. As a result, the speech depression detection task faces a notable challenge of data scarcity. Moreover, the issue of data imbalance is particularly acute in the dataset, as instances of healthy (positive cases) are significantly outnumbered by depression (negative) cases. We adopted Wu's approach (Wu et al., 2023) of augmenting the training set through sub-dialogue shuffling. Sub-dialogue shuffling involves sampling a sub-dialogue $x_{s:e}$ from each complete dialogue $x_{1:T}$, where s and e represent the randomly selected start and end utterance indexes, respectively.

This technique allowed us to balance the number of positive and negative samples effectively, while substantially increasing the dataset size. Differing from Wu's method, our use of landmarks in speech processing enables the use of longer sub-dialogues for training purposes. To ensure a fair comparison, we maintained the same data size (same sub-dialogue sampling number M=1000) as Wu's approach. For a detailed description of the algorithm, please refer to Appendix B.

Previous research has indicated that the patterns in which landmarks appear are more valuable than the individual landmarks themselves (Huang et al., 2019a). Therefore, as shown in Figure 1, we combined landmarks, treating every two consecutive landmarks as a single unit. This approach not only better represents the patterns of landmarks but also effectively reduces the length of the landmark sequence in each sample.

3.3 Hint Cross-modal Instruction Fine-Tuning

Since LLMs inherently lack exposure to acoustic landmarks, our initial step involves devising a method to teach the LLM what acoustic landmarks are. This foundational training is crucial for enabling the models to interpret and utilize acoustic landmark data effectively.

As depicted in the middle section of Figure 2, our

Method/ Model	Llama2-7B	Llama2-7B Chat	Llama2-13B	Llama2-13B Chat	GPT3.5	GPT4
Text Only	0.578	0.488	0.636	0.545	0.545	0.571
Landmark Only	0.521	0.434	0.559	0.538	-	-
Text + Landmark	0.545	0.500	0.695	0.666	-	-

Table 2: F1 scores for the different LLM models, We test all Llama2 models for 7B and 13B, also test on GPT.

task involves providing an LLM with instructions to predict potential acoustic landmarks based on text. This method serves a dual purpose: it enables the LLM to learn about acoustic landmarks, and it also aligns speech (landmarks) and text modalities using paired data. We adopt LoRA (Hu et al., 2022) by incorporating low-rank matrices into the Query and Key matrices of the self-attention layer, facilitating efficient adaptation and fine-tuning. Additionally, we resize the embedding layer of the LLMs to add the merged landmarks to the vocabulary. During the training process, both the **embedding layer**, **linear head** and the **LoRA matrices** are actively trained to integrate these new elements effectively. The training objective is to minimize the negative log-likelihood, and the loss calculation applies to all samples (including the prefix), which can be formulated as:

$$\mathcal{L}(M|C) = -\sum_{j=1}^{x}\sum_{i=1}^{y_j} \log P(s_{i,j}|s_{<i,j}, M), \quad (1)$$

where x is the number of samples in dataset C, y_j is the text and corresponding landmarks in sample S, and M denotes the large language model that we have fine-tuned.

Additionally, during dataset construction, we incorporate hints for the LLM. For example, when data are sourced from a patient with depression, we include a hint indicating their origin from a depressed patient. Experimentally, we found this method of data construction to be crucial, which also supports our hypothesis that **the acoustic landmarks from individuals with depression differ from those of healthy individuals**. For detailed template construction, please refer to Appendix C.

3.4 P-Tuning for Depression Detection

In the previous stage, we trained the LLMs to understand what landmarks are. Following this, we employ P-tuning (Liu et al., 2023) to enable the LLMs to integrate text and landmarks for depression detection. We replace the lm head layer with the classification layer. The training objective is to minimize cross-entropy for classification, which can be formulated as

$$\mathcal{L} = -\sum_{c=1}^{C} y_{o,c} \log(p_{o,c}), \quad (2)$$

where C is the number of classes. $y_{o,c}$ is an indicator variable that is 1 if the observation o belongs to class c and 0 otherwise. $p_{o,c}$ is the predicted probability of observation o belonging to class c. We also compared instruction tuning using LoRA with P-tuning and discovered that **manually constructed templates are not well-suited for depression classification tasks**. Furthermore, we observed a performance improvement when applying LoRA matrices across all layers of Llama2.

3.5 Decision Making

In the previous study by (Wu et al., 2023), they achieved state-of-the-art (SOTA) results through an ensemble approach, combining WavLM, WavLM pre-trained on emotional recognition tasks, and the combined result of RoBERTa and WavLM. Adopting a similar strategy, we fine-tune three distinct LlaMA2 (Text + Landmark) models, each with different data volumes (different numbers of sub-dialogue M(900, 1000, 1100)), and used them for ensemble voting.

4 Experiments

4.1 Experimental Setup

Dataset. The DAIC-WOZ dataset (DeVault et al., 2014), recognized as a standard for depression detection, includes 189 clinical interview recordings between interviewers and patients. In its training subset, 30 of the total 107 interviews are labelled as depressed, while the development subset contains 12 depressed instances out of 35 interviews. Consistently with previous studies (Gong and Poellabauer, 2017; Shen et al., 2022; Wu et al., 2022, 2023), we report our results on the development subset.

Model Configurations. Our research utilizes Llama2-7B, Llama-7B Chat, Llama2-13B, and Llama2-13B Chat, conducted on a system equipped with 8 NVIDIA A100 80GB GPUs. Llama 2-Chat was optimized for engaging in two-way conversations. In the cross-modal instruction fine-tuning

Methods	Model	F1	Ensemble
Previous SOTA (Wu et al., 2023)	WavLM + RoBERTa	0.648	0.829
	WavLM Layer 8	0.700	
	WavLM Layer 10	0.720	
Text+Landmark (Our)	Llama2 ($M=900$)	0.636	**0.833**
	Llama2 ($M=1000$)	0.695	
	Llama2 ($M=1100$)	0.719	

Table 3: A comparison of our proposed system with previous state-of-the-art (SOTA), where all ensemble outcomes(F1 Score) are derived from a majority vote. In the table, M denotes the number of augmented sub-dialogues per dialogue in our data augmentation algorithm, while the previous SOTA used M=1000 sub-dialogues.

stage, We fine-tuned the model with 10 epochs with 128 batch sizes, 8 Lora ranks, 100 warmup steps, and a 1e-6 learning rate. In the depression detection stage, we fine-tuned the model with 8 epochs with 256 batch sizes, 30 virtual tokens, 256 encoder hidden sizes, and a 1e-6 learning rate. In both experiments, we used AdamW as an optimizer with the model parallel to fine-tune our model. In the ablation study stage, we used hyperparameter tuning following the Tree-structured Parzen Estimator (TPE) paradigm (Bergstra et al., 2011).

4.2 Main Result: Performance of different LLMs in Depression Detection task

Depression Detection in Llama2. Table 2 displays the F1 scores obtained by Llama2 in depression detection across different scenarios. Additionally, we conducted a comparison of our findings with the results obtained from GPT-3.5 and GPT-4, focusing solely on their performance in the text modality. It is crucial to highlight that we did not fine-tune GPT-3 or GPT-4 for our purposes. Rather, we employed carefully crafted prompts(see appendix D), allowing the GPT models to assess whether a particular sample was from a patient with depression.

For the 'landmark only' and 'landmark + text' results, the process involved first undergoing hint cross-modal instruction fine-tuning and then employing P-tuning for depression detection. The objective was to equip the LLMs with a preliminary understanding of landmarks before advancing to the diagnostic stage for depression.

The experimental results reveal that when LLMs solely use the text modality for depression detection, the performance of all models, including notably powerful ones like GPT-3.5 and GPT-4, which excel in many tasks, is not particularly impressive and remains somewhat unsatisfactory. We attribute the subpar performance to two main factors. First is the **inherent limitation of the text modality in conveying emotional information**. For instance, consider the sentence, "It's raining today." While some may find this statement positive, others might feel the opposite. It's challenging to discern the emotional nuances from the text alone, but with audio information, we could accurately capture the emotional context of the statement. Secondly, **the issue lies with the data itself**. Labels are only available at the document level, and data are scarce (currently, there are no larger public datasets available for multimodal depression detection). This limitation in data granularity and volume significantly hinders the model's ability to accurately detect depression.

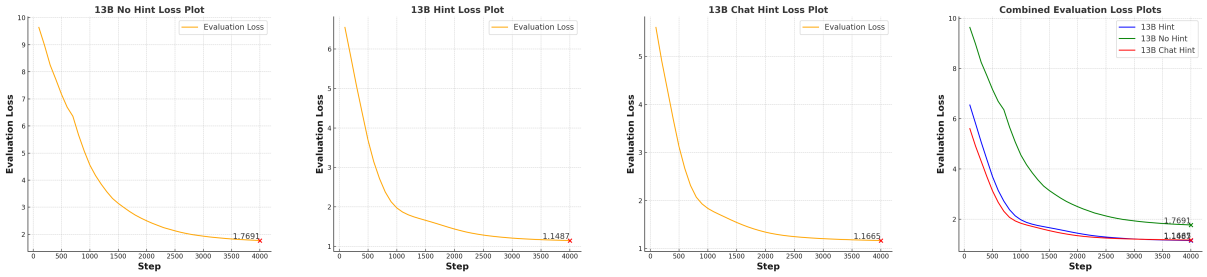
The introduction of landmarks led to enhanced performance across all models, affirming the effectiveness of our method in integrating landmarks. Landmarks can represent some of the acoustic information due to affective variation, providing additional information that assists LLMs in detecting depression. Nonetheless, the efficacy of using landmarks in isolation for depression detection was found to be suboptimal. Drawing on past research, we believe this is due to the fact that even after cross-modal instruction fine-tuning, relying solely on information from other modalities (such as audio or visual) could potentially impair the stability of LLMs (Zhang et al., 2023; Li et al., 2023c). When we combined multiple Llama2 models that had integrated both text and landmark information for depression detection, we achieved SOTA results as shown in table 3. Furthermore, as indicated in Table 3, there is a gradual improvement in Llama2's performance in depression detection tasks as the number of sub-dialogues per dialogue increases. This observation further emphasizes the crucial role that data quantity plays in the effectiveness of depression detection tasks.

5 Ablation Study and Discussion

In this chapter, we conduct an empirical study to meticulously analyze and elucidate the characteristics of LLMs that we identified in the context of depression detection during our experiments.

5.1 Effect of Hint in Cross-Modal Instruction Fine-Tuning

During the Cross-Modal Instruction Fine-Tuning phase, we discovered that providing a hint to the LLMs is crucial. In other words, informing the LLMs whether the data sample originates from a



(a) 13B No Hint (b) 13B Hint (c) 13B Chat Hint (d) Three Comparison

Figure 4: Evaluation loss for different configurations up to 4000 steps.

patient with depression significantly impacts the training outcome. As evident from Figure 4, without a hint, the loss converged to around 1.76 (as shown in Figure 4a). In contrast, with a hint, the loss consistently converged to near 1.1 (as depicted in Figures 4b and 4c). Figure 4d offers a more vivid illustration of the substantial difference that the presence or absence of a hint makes to the model's performance in our empirical study. This phenomenon supports our previous conjecture that **individuals with depression and those who are healthy differ in their vocal expressions and that landmarks are capable of reflecting this characteristic**. Although the differences between Llama2 and Llama2 Chat are not substantial, it is still observable that, in this phase, Llama2 outperforms its Chat version. We will provide a more detailed discussion in the subsequent section.

5.2 How LLMs Learn from Acoustic Landmarks

To further investigate how LLMs learn acoustic landmarks, we extended the application of LoRA beyond just the attention layers, applying it across all layers for comprehensive analysis (Pu et al., 2023; Sun et al., 2023; Li et al., 2023a; Zhang et al., 2024b). To find the matrix with the greatest contribution, we first need to define the method for calculating the contribution of a matrix. We can approximately consider the changes in the LoRA matrix as indicative of its contribution to the task (He et al., 2021). Therefore, we assess that the contribution of a matrix is calculated by summing the absolute values of all its elements, normalized by the total number of elements in the matrix. Suppose we have a set of LoRA matrices $L_1, L_2, \ldots, L_n$, each matrix L_i being an $a \times b$ matrix. Then, the contribution C_i of matrix L_i can be calculated using the formula:

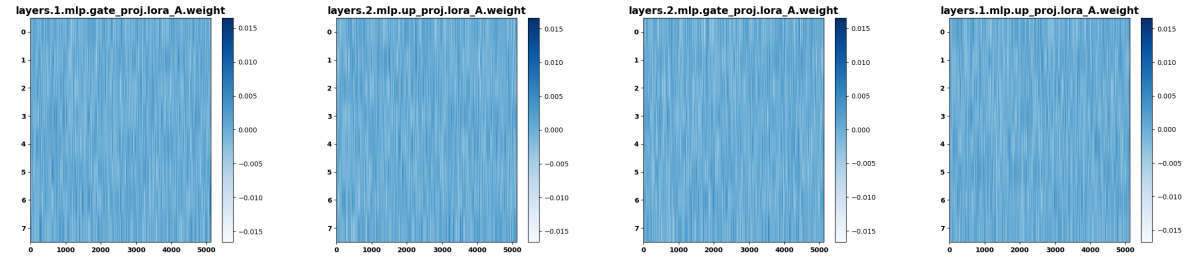
$$C_i = \frac{1}{ab} \sum_{j=1}^{a} \sum_{k=1}^{b} |L_i(j,k)|. \quad (3)$$

Here, $|L_i(j,k)|$ represents the absolute value of the element in the j^{th} row and k^{th} column of matrix L_i. After calculating the contribution value (C), we rank and select the ten matrices with the highest and the lowest contributions for further analysis. Figure 5 separately illustrates the four matrices with the greatest contributions and the four with the least. To validate the effectiveness of this method, we deactivated the five matrices with the smallest contributions and observed that this had no significant impact on our results.

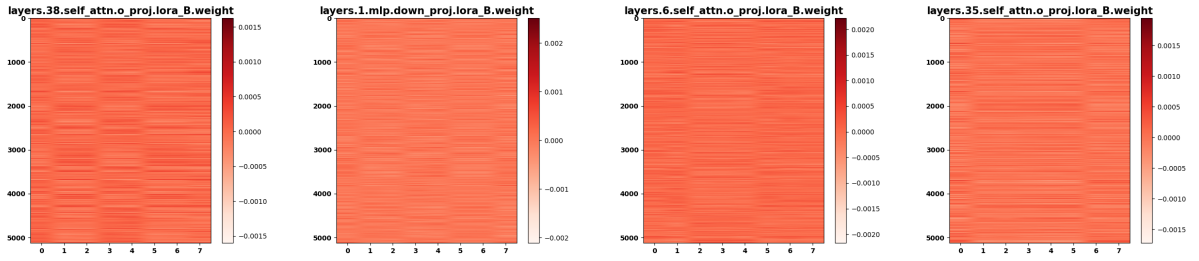
Our analysis of the matrices revealed that LLMs primarily **learn landmarks through the feedforward network**, while the contribution of the LoRA matrices in the attention layers is quite minimal. This phenomenon is also observed when training LLMs to learn speech codecs (Hao et al., 2023), suggesting that even though landmarks have inherent linguistic significance, LLMs tend to treat landmarks as abstract tensors, similar to speech codecs, during the learning process. Additionally, we observed that **layers closer to the beginning of the LLMs have a greater contribution** to learning landmarks. This could be because LLMs treat landmarks as new vocabulary items, leading to more updates in layers nearer to the embedding layer.

5.3 Llama2 vs Llama2 Chat, and Generation vs Classification

LlaMA2 models are uncensored and have not undergone instruction tuning or chat-tuning. In contrast, LlaMA2 Chat models are censored and have been chat-tuned, making them optimized for dialogue use cases (Touvron et al., 2023). When treating depression detection as a classification task,



(a) Top 1 Contribution Layer (b) Top 2 Contribution Layer (c) Top 3 Contribution Layer (d) Top 4 Contribution Layer



(e) Bottom Layer 1 (f) Bottom Layer 2 (g) Bottom Layer 3 (h) Bottom Layer 4

Figure 5: The top four images represent the LoRA matrices of the layers that contribute most significantly to the large language model's learning of landmarks. The bottom four images depict the LoRA matrices of the layers with the least contribution. As can be inferred from the graph's title, the feedforward layer is the primary contributor.

we tested LlaMA2 Chat and found that its performance, both during the Cross-modal Instruction Fine-Tuning stage and the depression detection phase, was inferior to that of LlaMA2. We hypothesize two potential reasons for this. The first is that the Chat version might not be suitable for classification tasks. The second, and our preferred explanation, is that the Chat version, having been adjusted, tends to avoid answering questions to mitigate ethical risks. To validate our hypothesis, we first reimagined the classification task as a generative task, where the LLMs diagnoses depression through dialogue responses. We tested this zero-shot scenario on GPT-3.5 and GPT-4. Additionally, we applied LoRA for instruction fine-tuning in various scenarios presented in Table 2, to observe how the models perform post-tuning. We observed that when treating depression detection as a generative task, neither LlaMA2 nor GPT models performed particularly well, with the dialogue-enhanced LlaMA Chat still underperforming compared with LlaMA. This suggests that LLMs in the field of depression detection are subject to certain artificial limitations, impacting their effectiveness in this specific application. The details of the template can be seen on Appendix D.

5.4 Lora VS P-tuning

From our previous ablation experiments, we found that the conventional method of incorporating LoRA matrices into attention layers might not be well-suited for depression detection tasks. After experimenting with applying LoRA matrices across all layers and conducting a hyperparameter search, we observed that LoRA, in this context, achieved results similar to those of P-tuning. Furthermore, in our use of LoRA for classification tasks, we tested a variety of manually crafted templates. However, none were as effective as using no task-specific prompt template. We believe this occurs because when we explicitly inform the LLMs that the task involves depression detection, the model tends to avoid responses that could pose ethical risks.

6 Conclusion

This paper introduces an efficient approach for depression detection using acoustic landmarks and LLMs. This approach is not only valuable for the detection of depression but also represents a new perspective in enhancing the ability of LLMs to comprehend speech signals. Furthermore, we are the **first to research multimodal depression detection using LLMs**. We establish a new benchmark with a SOTA F1-score of 0.84 through ensemble learning. Additionally, we evaluated various PEFT methods and discovered that applying Lora across all layers yields identical outcomes for both P-tuning and Lora in depression detection. Our analysis further reveals how LLMs process speech landmarks, guiding future research in this domain.

Limitations

In addition, The study is confined to the DAIC-WOZ dataset, which is currently the most commonly used and only publicly available dataset in the field of multimodal depression recognition, particularly in the area of speech. The difficulty in acquiring data due to numerous privacy concerns surrounding depression datasets is acknowledged. Despite the limitations of focusing on this single dataset, it aligns with traditional research methodologies in this domain, as previous studies have predominantly relied on it.

Ethics Statement

The DAIC-WOZ datasets are publicly available benchmarks and have been automatically de-identifed to protect patient privacy. Although our model improves the factual accuracy of generated reports, its performance still lags behind the needs of practical deployment. The outputs of our model may contain false observations and diagnoses due to systematic biases. In this regard, we strongly urge the users to examine the generated output in real-world applications cautiously.

Acknowledgement

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

References

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

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, et al. 2020. Language models are few-shot learners. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 33, pages 1877–1901.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

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.

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.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Hongkun Hao, Long Zhou, Shujie Liu, Jinyu Li, Shujie Hu, Rui Wang, and Furu Wei. 2023. Boosting large language model for speech synthesis: An empirical study. *arXiv preprint arXiv:2401.00246*.

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.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

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.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proc. Int. Conf. Learn. Representations*.

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.

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.

Shuyue Stella Li, Beining Xu, Xiangyu Zhang, Hexin Liu, Wenhan Chao, and Paola Garcia. 2023a. A quantitative approach to understand self-supervised models as cross-lingual feature extracters. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 200–211.

Shuyue Stella Li, Xiangyu Zhang, Shu Zhou, Hongchao Shu, Ruixing Liang, Hexin Liu, and Leibny Paola Garcia. 2023b. Pqlm-multilingual decentralized portable quantum language model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023c. Prompting large language models for zero-shot domain adaptation in speech recognition. *arXiv preprint arXiv:2306.16007*.

Hexin Liu, Leibny Paola Garcia Perera, 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, Leibny Paola Garcia, Andy WH 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.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

William S Noble. 2006. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.

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.

George Pu, Anirudh Jain, Jihan Yin, and Russell Kaplan. 2023. Empirical analysis of the strengths and weaknesses of peft techniques for llms. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

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.

Kenneth N Stevens. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4):1872–1891.

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*.

Xianghui Sun, Yunjie Ji, Baochang Ma, and Xiangang Li. 2023. A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *arXiv preprint arXiv:2304.08109*.

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*.

Jane Walker, Katy Burke, Marta Wanat, Rebecca Fisher, Josephine Fielding, Amy Mulick, Stephen Puntis, Joseph Sharpe, Michelle Degli Esposti, Eli Harriss, et al. 2018. The prevalence of depression in general

hospital inpatients: a systematic review and meta-analysis of interview-based studies. *Psychological medicine*, 48(14):2285–2298.

Zhuo Wang, Rongzhen Li, Bowen Dong, Jie Wang, Xiuxing Li, Ning Liu, Chenhui Mao, Wei Zhang, Liling Dong, Jing Gao, et al. 2023. Can llms like gpt-4 outperform traditional ai tools in dementia diagnosis? maybe, but not today. *arXiv preprint arXiv:2306.01499*.

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. 2023. 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.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

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

Xiangyu Zhang, Jianbo Ma, Mostafa Shahin, Beena Ahmed, and Julien Epps. 2024b. Rethinking mamba in speech processing by self-supervised models. *arXiv preprint arXiv:2409.07273*.

Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. 2024c. Mamba in speech: Towards an alternative to self-attention. *arXiv preprint arXiv:2405.12609*.

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.

Wenbo Zheng, Lan Yan, and Fei-Yue Wang. 2023. Two birds with one stone: Knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. *IEEE Transactions on Affective Computing*.

A Details of Landmark Detection

A.1 General Processing Details

Given a discrete time series signal $x[n]$, the process of peak detection consists of several pre-processing steps, followed by the identification of significant peaks. The steps are as follows:

Six Frequency Bands

The following table describes the six frequency bands we used in our algorithm.

Table 4: Frequency Bands

Band	Frequency Range (kHz)
1	0.0–0.4
2	0.8–1.5
3	1.2–2.0
4	2.0–3.5
5	3.5–5.0
6	5.0–8.0

Coarse Smoothing

The signal is first subjected to a coarse smoothing operation to reduce noise and highlight broader trends. This is achieved by applying a centered moving average with a window size of cp_sm:

$$L_b^{(cp)}[n] = 10 \cdot \log_{10}\left(\frac{1}{2N_{cp}+1}\sum_{k=-N_{cp}}^{N_{cp}} E_b[n+k]\right) \quad (4)$$

where $E_b[n]$ is the energy in the b^{th} frequency band at time n, and N_{cp} is half the size of the coarse smoothing window.

Coarse Differentiation

The smoothed signal undergoes differentiation to identify regions of rapid change, which could indicate potential peaks. The differentiation is centered on mitigating delay:

$$D_b^{(cp)}[n] = L_b^{(cp)}[n + cp\_dt] - L_b^{(cp)}[n], \quad (5)$$

followed by a shift to center the result:

$$D_b^{(cp)}[n] \leftarrow D_b^{(cp)}[n - \lfloor cp\_dt/2 \rfloor]. \quad (6)$$

Fine Smoothing

A finer smoothing operation is applied to the original signal to preserve more detail, with a window size of fp_sm:

$$L_b^{(fp)}[n] = 10 \cdot \log_{10}\left(\frac{1}{2N_{fp}+1}\sum_{k=-N_{fp}}^{N_{fp}} E_b[n+k]\right) \quad (7)$$

where N_{fp} is half the size of the fine smoothing window.

Fine Differentiation

As with coarse differentiation, the finely smoothed signal is differentiated:

$$D_b^{(fp)}[n] = L_b^{(fp)}[n + fp\_dt] - L_b^{(fp)}[n], \quad (8)$$

and then centered:

$$D_b^{(fp)}[n] \leftarrow D_b^{(fp)}[n - \lfloor fp\_dt/2 \rfloor]. \quad (9)$$

Peak Detection

After pre-processing, peaks are identified using the conditions specified earlier, considering factors such as prominence, height, and minimum distance between peaks.

Given a signal sequence $x[n]$, the peak detection process can be mathematically described as follows:

A data point $x[n]$ is considered a local maximum if it satisfies the following condition:

$$x[n] > x[n-1] \quad \text{and} \quad x[n] > x[n+1]. \quad (10)$$

If a height threshold h is specified, $x[i]$ is recognized as a peak only if:

$$x[i] > h. \quad (11)$$

The prominence P of a peak at $x[i]$ is defined as the vertical distance between the peak and its lowest contour line:

$$P = x[i] - \max(v_l, v_r), \quad (12)$$

where v_l and v_r are the lowest points on either side of $x[i]$, before reaching a higher point. A peak is considered significant if its prominence exceeds a predefined threshold.

The width W of a peak is measured at a vertical distance P from its highest point. Points $x[l]$ and $x[r]$, where $l < i < r$, are the positions at which the signal drops below the threshold defined by the prominence:

$$x[l] < x[i] - P \quad \text{and} \quad x[r] < x[i] - P, \quad (13)$$

and the width W is the distance between $x[l]$ and $x[r]$.

If a minimum peak separation distance D is defined, then for any two peaks $x[i]$ and $x[j]$, the condition must be met:

$$|i - j| > D. \quad (14)$$

These conditions are used to identify peaks in the signal that are not only local maxima but also exceed certain amplitude and prominence thresholds, ensuring the detected peaks are significant in the context of the signal.

A.2 Details of Specific Landmark Detection

g landmark When both the coarse and fine filters exhibit a peak in band 1, it is identified as a 'g' landmark.

b landmark In an unvoiced segment (not between +g and the next -g), if at least three out of five frequency bands demonstrate simultaneous power increases of no less than 6 dB in both coarse and fine filters, a specific condition or criterion is met.

s landmark In an unvoiced segment (between +g and the next -g), if at least three out of five frequency bands demonstrate simultaneous power increases of no less than 6 dB in both coarse and fine filters, a specific condition or criterion is met.

f+ and **v+ landmarks** involves detecting a 6 dB power increase in at least three high-frequency bands (4, 5, 6), and a power decrease in low-frequency bands (2, 3). For **f-** and **v-**, the criteria are reversed: a 6 dB power decrease in the same high-frequency bands and a power increase in the low-frequency bands.The distinguishing factor here is that frication landmarks are detected within unvoiced segments (b landmark), while voiced frication landmarks are sought in voiced segments (s landmark).

p landmark, p landmark extraction can be divided into several steps.

1. **Frame Segmentation**:
Let the audio signal be $Y(t)$.
Define the frame length N and frame shift Δ.
For the i-th frame, we consider the segment $Y[i \cdot \Delta : i \cdot \Delta + N]$.

2. **Autocorrelation Calculation**:
For each frame Y_i, calculate the autocorrelation function $R_{xx}(k)$:

$$R_{xx}(k) = \frac{1}{N-k} \sum_{n=0}^{N-k-1} Y_i(n) \cdot Y_i(n+k).$$

3. **Energy Function Calculation**:
Compute the energy function E_f for each frame:

$$E_f(i) = \frac{1}{N} \sum_{k=0}^{N-1} R_{xx}(k)^2.$$

4. **Upsampling**:
Upsample the energy function E_f to match the length of the original signal.

5. **Smoothing**:

Algorithm 1 Sub-dialogue shuffling

1: $N^+ \leftarrow$ Number of positive samples in the training set
2: $N^- \leftarrow$ Number of negative samples in the training set
3: $M \leftarrow$ Set number of sub-dialogues for each positive sample M^+
4: $M^* \leftarrow N^-/N^+$
5: Set $\varepsilon_l, \varepsilon_h$ satisfying $0 < \varepsilon_l < \varepsilon_h \leq 1$
6: **for** Dialogue $X^{(n)}$ $n = 1$ **to** N **do**
7: $T \leftarrow \text{len}(x^{(n)})$
8: **if** $x^{(n)}$ is positive **then**
9: $M \leftarrow M^+$
10: **else**
11: $M \leftarrow M^-$
12: **end if**
13: **for** Sub-dialogue $X^{(n)m}$ $m = 1$ **to** M **do**
14: Sample ε uniformly from $(\varepsilon_l, \varepsilon_h)$
15: $d \leftarrow \varepsilon T - 1$
16: Sample s randomly from range $(0, T - d)$
17: $e \leftarrow s + d$
18: $X^{(n)m} \leftarrow x^{(n)}_{s:e}$
19: **end for**
20: **end for**

Apply smoothing(As defined in the previous section) to the upsampled energy function.
6. **Binarization**:
Define a threshold θ, and convert the smoothed energy function into a binary signal.
7. **Jump Detection**:
Detect positive and negative jumps in the binary signal.
8. **P Landmark Index and Time Determination**:
Record the positions of jumps, which are the indices of P landmarks.
Convert these indices into time points to determine the P landmarks.

B Details of Data Augmentation

The training set was expanded by shuffling sub-dialogues, selecting portions $x_{s:e}$ from each full dialogue $x_{1:T}$, with s and e as random start and end indices. The algorithm outlines this process. Initially, it counts the positive and negative samples, setting M^+ as the target number of sub-dialogues for each positive dialogue (Algorithm 1, lines 1-3). To balance augmentation, M^- is calculated using N^+, N^-, and M^+ (line 4). For both positive and negative dialogues, corresponding M^+ and M^- sub-dialogues are generated (lines 8-12). The sub-dialogue length, d, is set within the range defined by ε_l and ε_h, chosen randomly (lines 14-15). The start index s is randomly selected within its range, and the end index e is determined accordingly (lines 16-18) (Wu et al., 2023).

C Sample of Hint Cross-modal Instruction Fine Tuning

Depression Example

```
Below are the speech transcripts from a
    person with depression.
Please try to predict the concatenated
    acoustic landmarks
corresponding to these transcripts.

### Transcript:
{transcript}

### Acoustic Landmark:
{landmark}
```

Healthy Example

```
Below are the speech transcripts from a
    healthy person.
Please try to predict the concatenated
    acoustic landmarks
corresponding to these transcripts.

### Transcript:
{transcript}

### Acoustic Landmark:
{landmark}
```

D Sample of Instruction Fine-Tuning for Depression Detection

Text Only

```
"Categorize these dialogues as either
    depression or healthy based on its
    transcripts.

### transcript:{transcript}

### Response:"
```

Landmark Only

```
"Categorize these dialogues as either
    depression or healthy based on its
    acoustic landmarks.

### acoustic landmarks:{landmarks}

### Response:"
```

MultiModal

```
"Categorize these dialogues as either
    depression or healthy based on its
    transcripts and acoustic landmarks.

### Transcript:{transcript}

### Acoustic Landmark:{landmarks}

### Response:\n"
```