Beyond Silent Letters: Amplifying LLMs in Emotion Recognition with Vocal Nuances

Anonymous ACL submission

Abstract

Emotion recognition in speech is a challenging multimodal task that requires understanding both verbal content and vocal nuances. This paper introduces a novel approach to emotion detection using Large Language Models (LLMs), which have demonstrated exceptional capabilities in natural language understanding. To overcome the inherent limitation of LLMs in processing audio inputs, we propose *SpeechCueLLM*, a method that translates speech characteristics into natural language descriptions, allowing LLMs to perform multimodal emotion analysis via text prompts without any architectural changes. Our method is minimal yet impactful, outperforming baseline models that require structural modifications. We evaluate *SpeechCueLLM* on two datasets: IEMOCAP and MELD, showing significant improvements in emotion recognition accuracy, particularly for high-quality audio data. We also explore the effectiveness of various feature representations and fine-tuning strategies for different LLMs. Our experiments demonstrate that incorporating speech descriptions yields a more than 2% increase in the average weighted F1 score on IEMOCAP (from 70.111% to 72.596%). The source code will be made available on GitHub.

1 Introduction

Emotion detection in speech is a crucial aspect of human-computer interaction, with applications spanning from customer service to mental health monitoring (Van Thanh, 2023; Dhanasekar et al., 2021). Recent research has demonstrated that Large Language Models (LLMs) possess a form of emotional intelligence, capable of interpreting emotional stimuli in text (Li et al., 2023a). Advances in Emotion Recognition in Conversation (ERC) have utilized LLMs to improve performance and generalizability (Lei et al., 2024; Xue et al., 2024; Fu, 2024).

However, despite the remarkable progress of LLMs in handling textual data, they remain fundamentally limited in processing audio inputs. This limitation constrains their utility in multimodal emotion recognition tasks, which require the integration of both verbal content and vocal nuances. While recent studies have explored incorporating speech features into LLM-based systems (Xu et al., 2023; Zhang et al., 2024a; Gong et al., 2023; Santoso et al., 2024), these approaches often introduce additional model parameters and pre-training stages, increasing complexity and computational overhead.

In this paper, we address these challenges by proposing *SpeechCueLLM*, a novel method that translates speech characteristics into natural language descriptions, enabling LLMs to process multimodal data through text prompts without requiring any architectural changes. This approach bridges the gap between audio and text modalities, allowing LLMs to conduct multimodal emotion analysis. *SpeechCueLLM* sets itself apart from existing speech-integrated LLM approaches in three key aspects: **(a) Direct Integration**: Our method allows seamless integration of speech characteristics without adding new neural components (e.g., audio encoders or projection layers), thus simplifying the training process. **(b) End-to-End Fine-tuning**: We fine-tune the LLMs directly on inputs that combine textual data with speech descriptions, eliminating the need for weak label generation or additional training stages. **(c) Flexibility and Generalization**: By translating speech features into natural language descriptions, *SpeechCueLLM* provides a flexible, plug-in solution for incorporating speech information into LLM-based systems without modifying the LLM architecture.

Our main contributions are as follows:

1. We introduce *SpeechCueLLM*, a method that enables multimodal analysis by integrating

speech characteristics into text prompts, facilitating emotion recognition without architectural changes to the LLMs. *SpeechCueLLM* outperforms traditional approaches that incorporate speech encoders for embedding audio features into LLMs.

2. We present key findings on the effectiveness and limitations of this method, showing that incorporating speech descriptions enhances emotion recognition accuracy across multiple LLMs while offering insights into the role of audio quality and the advantages of prompt-based techniques.

This study aims to showcase the potential of *SpeechCueLLM* to enhance the emotion recognition capabilities of LLMs while providing insights into the factors driving its effectiveness. We evaluate *SpeechCueLLM* on two widely used datasets in ERC: IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019) (Section 4.1), achieving state-of-the-art performance. We compare our method with projection-based approaches that incorporate a separate speech encoder and show that *SpeechCueLLM* surpasses these baseline methods (Section 4.3). Additionally, we explore the impact of different speech feature representations (Section 4.4.1) and examine the fine-tuning performance of various LLMs (Section 4.4.3).

2 Related Work

2.1 Multimodal Emotion Models

Exsiting multimodal methods have been proposed to combine audio and text modalities for emotion recognition (Zadeh et al., 2017; Wu et al., 2024; Meng et al., 2024; Tsai et al., 2019; Goncalves and Busso, 2022; Ma et al., 2024a). Many of these works utilize transformer encoder structures to fuse multimodal signals through cross-modality attention mechanisms (Tsai et al., 2019). Additionally, some models employ variations of Graph Neural Networks (GNNs) to capture multimodal semantic features and enhance emotion recognition accuracy (Meng et al., 2024). These methods focus on model structure that captures and learns from multimodal signals holistically.

2.2 LLMs for Emotion Recognition

Recent advancements in ERC have leveraged LLMs to enhance performance and generalizability. In our work, we build upon the InstructERC framework (Lei et al., 2024), which reformulates emotion recognition as a generative task using LLMs. While InstructERC introduced retrieval templates and emotion alignment tasks, recent extensions such as BiosERC (Xue et al., 2024), CKERC (Fu, 2024), and DialogueLLM (Zhang et al., 2024b) have incorporated additional contextual information like biographical data, commonsense knowledge, and visual descriptions.

Our approach distinguishes itself by focusing on integrating speech characteristics into the templates as natural language descriptions, bridging the gap between audio and text modalities in LLM-based emotion recognition systems.

2.3 Speech Incorporation in LLMs

Recent works have explored integrating speech features into LLM-based systems for various downstream tasks. SECap (Xu et al., 2023) uses LLaMA as a text decoder to generate speech emotion captions, relying on an additional audio encoder and bridging network. Similarly, (Zhang et al., 2024a) incorporates acoustic landmark tokens for depression detection and requires extra parameters and pre-training.

LanSER (Gong et al., 2023) infers weak emotion labels from speech transcripts to use unlabeled data, while (Santoso et al., 2024) focuses on emotional speech annotation, combining conversation context with acoustic feature descriptors. While this approach shares our interest in combining textual and acoustic information in the LLM inputs, it primarily uses LLMs for weak label generation rather than direct emotion recognition training.

Recent research has also explored the application of LLMs in Automatic Speech Recognition (ASR)(Bai et al., 2024; Baskar et al., 2024; Ma et al., 2024b) , demonstrating the versatility of LLMs in speech-processing tasks. Further extending the capabilities of LLMs in speech processing, AudioChatLlama (Fathullah et al., 2024) and Audio-LLM (Li et al., 2024) presents end-to-end model that integrates general-purpose speech abilities into the Llama-2 model.

SpeechCueLLM differs from these approaches by processing audio-derived information without requiring architectural modifications. This method is highly compatible with existing LLM infrastructures, facilitating easy implementation and integration into current systems.

3 Methodology

SpeechCueLLM integrates descriptions of speech characteristics in natural language into the prompt for LLMs, as illustrated in Figure 1. Traditional LLMs, when limited to textual inputs, omit critical emotional cues that are embedded in audio signals. This omission can result in a failure to capture important aspects of emotion, as the same sentence can express different emotions based on variations in pitch, volume, intonation, and other vocal features. By translating audio signals into natural language descriptions, *SpeechCueLLM* enables LLMs to perform more accurate and nuanced multimodal emotion recognition.

3.1 Audio Features

In our approach to emotion detection, we leverage five well-established and intuitive audio features that convey emotional cues (Koolagudi and Rao, 2012; Alva et al., 2015; Sudhakar and Anil, 2015). These features are designed to be easily interpreted by both humans and LLMs, capturing key aspects of speech that reflect emotional information.

The first feature, *volume*, is represented by two sub-features: *average volume*, which indicates the overall loudness of the speech, and *volume variation*, which captures dynamic changes in intensity. These features allow us to detect emotional nuances based on loudness.

Pitch, another critical indicator of emotion, is also divided into two sub-features: *average pitch*, reflecting the general tone of the speaker's voice, and *pitch variation*, which highlights modulation and changes in pitch, providing deeper insights into the speaker's emotional state.

The fifth feature, *speaking rate*, measures the speed of speech, offering valuable information about the speaker's emotional urgency or level of excitement.

By focusing on these fundamental yet comprehensive audio characteristics, we ensure that the extracted features are rich in emotional content while remaining interpretable by LLMs when translated into natural language descriptions.

3.1.1 Audio Feature Processing

Our approach to audio feature processing transforms raw numerical features into categorical representations to enhance interpretability and facilitate natural language descriptions. This process is composed of four key steps:

Description:
Target speech characteristics: moderate pitch with high variation, high volume with moderate variation, very low speaking rate.

Impression:
The target speaker has a moderate pitch with likely noticeable variation, suggesting expressiveness, while likely speaking loudly, which might indicate excitement, confidence, or urgency, with normal volume variation, and is talking slowly, possibly suggesting thoughtfulness, hesitation, or calmness.

Figure 1: Example of speech characteristic description (top) and derived impression (bottom). Each color represents the same set of features (■ pitch, ■ volume, and ■ speaking rate).

1. **Threshold Calculation**: We compute thresholds for each audio feature (such as average volume, pitch variation, and speaking rate) using quantile-based segmentation. Depending on the desired granularity, the feature space is divided into 3, 4, 5, or 6 classes, with thresholds determined by appropriate quantiles. Detailed quantile splits can be found in Table 8 in Appendix A.3.

2. **Speaker-Specific Normalization**: To account for individual differences in speech patterns, we calculate speaker-specific thresholds, falling back to overall thresholds for less frequent speakers. This allows for more accurate categorization of speech features relative to each speaker's typical speaking style.

3. **Categorization**: Based on the calculated thresholds, each numerical feature is categorized. For example, in a 5-class system, a feature value may be classified as "very low," "low," "medium," "high," or "very high." This categorization provides a more intuitive, human-readable representation of the audio features.

4. **Feature-Specific Descriptions**: For each key audio feature (volume, pitch, speaking rate), we generate descriptive phrases corresponding to their categorical values. For instance, "high volume with moderate variation" or "low pitch with high variation" offers a more accessible and interpretable description of the audio features.

3.1.2 Audio Impression

To make the audio features more accessible and meaningful for both human interpretation and LLM

processing, we generate hard-coded natural language impressions based on the categorized features. These impressions go beyond simple descriptions, offering interpretative insights that suggest potential emotional or psychological states associated with the observed speech patterns. The feature-impression mapping is manually curated based on established relationships between prosodic features and emotional states (Mozziconacci, 2002; Frick, 1985).

For example, "speaking loudly with significant volume changes" might be interpreted as indicating "excitement, confidence, or urgency." The full mapping of features to impressions can be found in Table 9 in Appendix A.3.

To account for the inherent uncertainty in interpreting speech patterns, we incorporate hedge words (e.g., "likely," "may") depending on how close a feature is to its categorical threshold. This nuanced approach prevents overconfident interpretations, particularly in borderline cases.

The final impression synthesizes information from pitch, volume, and speaking rate into a coherent, flowing sentence. This integration provides a rich description of the speaker's vocal characteristics and their potential emotional implications.

3.2 *SpeechCueLLM* Prompt Template

Our methodology employs a comprehensive prompt template designed to enhance emotion detection performance. As shown in Figure 2, the template is structured around four key components:

1. **Instruction**: This component sets the stage by positioning the LLM as an expert in sentiment and emotion analysis, priming it for the task.
2. **Context**: Providing conversational background is essential for helping the LLM understand the broader situation and interpersonal dynamics. Additionally, speech features are appended to the last three utterances, offering supplementary speech context. More details on the added speech context are discussed in Appendix A.1.
3. **Speech Descriptions**: This innovative element translates audio signals into natural language descriptions of key speech characteristics like volume, pitch, and speaking rate. By including these paralinguistic details, we bridge the gap between audio and text, enabling the LLM to account for speech features that are crucial for accurate emotion detection.
4. **Question**: This final component asks the LLM to select an emotional label for a target utterance from a predefined set of options. It explicitly directs the model to consider both the contextual background and the described audio features in making its decision.

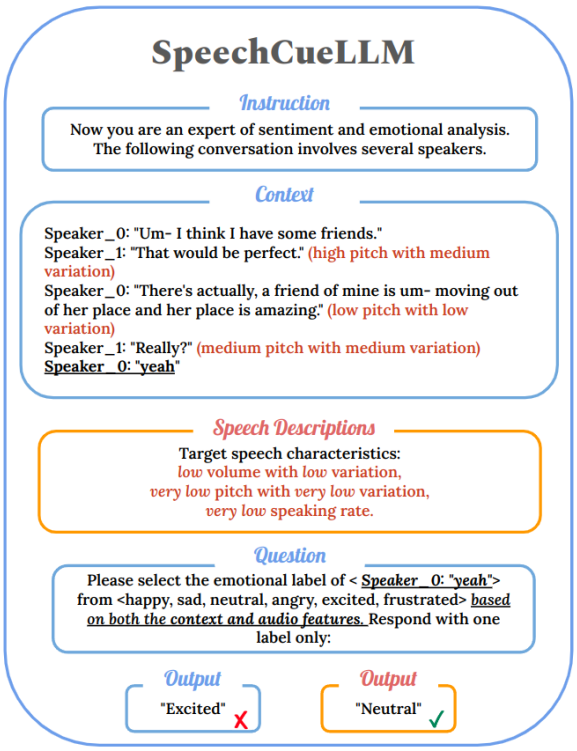

Figure 2: *SpeechCueLLM*Prompt Template for Emotion Detection: the last bold sentence with an underline represents the target utterance. The orange section highlights the outputs with added speech descriptions. This structured template integrates both textual context and speech characteristics to guide the LLM in performing multimodal emotion analysis.

Our empirical analysis shows that straightforward prompts with simple instructions yield optimal results. This structured approach ensures that the LLM can access both textual and audio-derived information, leading to a more nuanced and accurate emotion detection process. By incorporating speech characteristics into the prompt, we allow the LLM to perform multimodal analysis without requiring direct audio input processing capabilities.

4 Experiments

4.1 Datasets

We utilize two widely recognized datasets in the field of multimodal emotion recognition:

IEMOCAP (Interactive Emotional Dyadic Motion Capture) (Busso et al., 2008): IEMOCAP consists of dyadic conversations between actors,

providing a rich set of emotional expressions in a controlled environment. It contains approximately 12 hours of audiovisual data from 10 actors, divided into 5 sessions. Each utterance is annotated with one of nine emotion labels: anger, happiness, excitement, sadness, frustration, fear, surprise, neutral, and other. IEMOCAP is particularly valued for its high-quality audio recordings and motion capture data, making it an ideal dataset for multimodal emotion analysis.

MELD (Multimodal EmotionLines Dataset) (Poria et al., 2019): MELD is derived from the TV series *Friends*, making it particularly useful for analyzing emotions in natural conversational settings. It contains 13,708 utterances from 1,433 dialogues, each annotated with one of seven emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise.

In our experiments, we focus on the audio modality from both datasets. For the MELD dataset, we implement speaker-based standardization of audio features to account for individual vocal characteristics. In the IEMOCAP dataset, we apply gender-based standardization to account for physiological differences that can affect vocal features. Additionally, we confine our analysis of IEMOCAP to six emotion categories (anger, happiness, excitement, sadness, frustration, and neutral), following common practices in the field (Gong et al., 2024).

Our use of both the MELD and IEMOCAP datasets is designed to evaluate our approach across diverse audio environments. MELD, originating from a television show, presents more chaotic and naturalistic scenarios, while IEMOCAP offers a more controlled, studio-based environment. This dual-dataset approach enables a comprehensive assessment of our emotion detection method under varied recording conditions.

4.2 Experiment Settings

4.2.1 *SpeechCueLLM* Training Settings

For our experiments, we focus on open-sourced LLMs and include comparisons with the best-performing proprietary LLMs. We input the structured prompts into the LLMs and apply LoRA (Hu et al., 2021), an efficient low-parameter fine-tuning method that minimizes overfitting. This method trains the model to output a single emotion label based on the prompt. To further validate our approach, we also conduct experiments using only speech features in the prompts, demonstrating the effectiveness of these simple features.

We experiment with fine-tuning various LLMs, including LLaMA-2, multiple versions of LLaMA-3, and phi-3-medium, to evaluate the generalizability of our approach. For training, we use a learning rate of 3e-4 for IEMOCAP and 2e-4 for MELD, with a linear warm-up scheduler. Additionally, we test the zero-shot performance of Claude Sonnet 3.5 for comparison. Results are averaged over three independent runs to ensure robustness.

4.2.2 Baselines: LLMs with Speech Encoder

To benchmark our approach, we implement a projection-based method that leverages speech features by training additional parameters. In this setup, audio features are encoded using a pre-trained Automatic Speech Recognition (ASR) model and projected into the LLM's embedding space, as illustrated in Figure 4. This approach is inspired by SLAM-ASR (Ma et al., 2024b), which combines off-the-shelf speech encoders with LLMs using a trainable linear projector, achieving state-of-the-art performance on the Librispeech benchmark (Panayotov et al., 2015).

For consistency, the text portion of the input follows the same prompt format as in our main approach, but without the inclusion of speech-derived features. The encoded audio tokens are placed before the text tokens in the input sequence.

We use the Data2Vec model (Baevski et al., 2022) as the speech encoder. To prevent overfitting, we apply a downsampler to reduce the feature sequence length from 50 tokens per second to 10 tokens per second. For projecting the speech features into the LLM embedding space, we employ two types of projectors:

1. A straightforward feed-forward neural network.
2. A Q-former (Li et al., 2023b), a query-based transformer model that uses learnable queries to attend to speech features, producing a condensed and structured representation of key information from the audio.

Both projectors transform the encoded audio features into a format compatible with the LLM's token embeddings. As with the main approach, we fine-tune the LLM backbone using LoRA for all baseline methods.

4.3 Results

Our results show that *SpeechCueLLM* outperforms baselines that embed audio features using speech

Approach	Trainable parameters	IEMOCAP F1 Score	MELD F1 Score
Speech-Encoder-Only	93.2M	52.312 $\pm$1.7 (48.263)	47.921 $\pm$0.27 (28.413)
LLM (w/o Speech Features)	12.6M	70.111 $\pm$0.36	67.44 $\pm$0.26
+FeedForward$_{\text{Projector}}$ +SE$_{\text{Freeze}}$	28.9M	70.221$\pm$0.2	67.402$\pm$0.27
+FeedForward$_{\text{Projector}}$ +SE$_{\text{Unfreeze}}$	122.1M	70.635$\pm$0.74	66.156$\pm$0.83
+Q-Former$_{\text{Projector}}$ +SE$_{\text{Freeze}}$	71.4M	70.235$\pm$0.66	67.194$\pm$0.15
SpeechCueLLM(Ours)	12.6M	**72.596** $\pm$0.26	**67.604** $\pm$0.38

Table 1: Performance (weighted F1) comparison between *SpeechCueLLM*and the baselines. The top section includes results using only the speech encoder. (The number in parenthesis is the average F1 score.) The middle section includes results using only the LLM without speech features and varitions of the projection-based model. The bottom is our *SpeechCueLLM*results.

encoders, achieving state-of-the-art (SOTA) performance, as shown in Table 1.

We evaluate the effectiveness of speech features fine-tuned with the speech encoder. When used standalone, the audio encoder achieves a weighted F1 score of around 50 on both datasets. However, for MELD, the model tends to overfit the majority class, resulting in a reduced average F1 score of 28.413, while IEMOCAP achieves 48.263. This disparity in performance between the two datasets can be attributed to differences in audio quality and recording conditions. IEMOCAP, recorded in a controlled studio environment, provides clean and consistent audio, facilitating more accurate speech feature extraction. In contrast, MELD, derived from a TV show, *Friends*, presents more challenging audio conditions. The presence of background laughter, varying audio quality, overlapping speakers, and short utterance durations (averaging 2-3 seconds) likely contribute to less reliable speech feature extraction, thereby limiting the benefits of speech features.

We conduct several experiments with different variations, including alternative projectors and either freezing or training the speech encoder. The results, shown in Table 1, indicate slight improvements for IEMOCAP over the text-only model across all variations. Unfreezing the speech encoder leads to a marginally higher average score of 70.635%, though at the cost of increased performance variability. Notably, the Q-former, despite having significantly more parameters than the feed-forward layers, delivers comparable results. In contrast, for MELD, all variations perform worse than the text-only model, likely due to the poor audio quality.

The performance gains in IEMOCAP from the projection-based approach (0.1-0.5) are notably smaller than those achieved using the prompt-based method (2.4). The prompt-based approach consistently outperforms the direct integration of speech encoder features into the LLM embedding space for several reasons. First, it preserves contextual coherence by representing speech characteristics as natural language descriptions, allowing the LLM to leverage its strengths in language processing. In contrast, inserting speech embeddings into the LLM's embedding space can cause discrepancies in the distribution, introducing training complexity. This method may benefit from additional pre-training to align embeddings from the two modalities using large amount of text-audio paired data, as supported by previous research (Xu et al., 2023; Li et al., 2019; Arjmand et al., 2021). By sidestepping these multimodal alignment challenges, the prompt-based method delivers more consistent and reliable results without requiring extra tunings.

Furthermore, the prompt-based method enhances generalization by converting abstract speech features into textual descriptions, making it adaptable to various tasks and languages. This approach also offers greater interpretability, as the model explicitly receives human-readable descriptions of vocal nuances, unlike embedding modifications that obscure the model's processing of audio features. Finally, the prompt-based method is computationally efficient, avoiding the need for complex architectural modifications or additional training, making it a lightweight yet highly effective solution.

4.4 Ablation Study and Discussion

4.4.1 How Effective Are Various Speech Features?

The results in Table 2 demonstrate that incorporating speech features into text input improves performance across both datasets, though to varying degrees. For the IEMOCAP dataset, we observe a significant improvement of nearly 2 percentage

Method	Text-only	+ speech des	+ speech imp	+ speech des with context
IEMOCAP	70.111 ±0.36	72.021 ±0.54 (+1.910)	71.542 ±1.12 (+1.431)	**72.596** ±0.26 (+2.485)
MELD	67.44 ±0.26	67.074 ±0.49 (-0.366)	**67.604** ±0.38 (+0.164)	67.09 ±0.57 (-0.35)

Table 2: *SpeechCueLLM*performances (weighted F1 score in %) on IEMOCAP and MELD datasets using LoRA fine-tuning on the LLaMA-2-7B-base model. Results are averaged over three independent runs, reported as mean ± standard deviation. Dark green values in parentheses show the improvement over the Text-only method.

points when adding speech descriptions to the text input (from 70.111% to 72.021%). By further enriching the input with additional speech features for context, the model improves to 72.596% (detailed context experiments are available in Appendix A.1). This notable gain highlights the importance of integrating acoustic information in emotion recognition tasks, particularly when high-quality audio recordings are available.

In contrast, the MELD dataset shows no improvement when using speech descriptions and only a modest improvement when adding speech impressions, with a slight increase from 67.44% to 67.604%. This mirrors the trends observed in the projection-based model experiments, where the lower audio quality in MELD limits the effectiveness of speech feature integration.

Interestingly, the inclusion of speech impressions does not result in further improvements over speech descriptions for either dataset. In IEMOCAP, the performance (71.542%) with speech impressions falls between the text-only and text with speech descriptions results, while in MELD, it performs similarly to the text-only baseline (67.604%). This suggests that the hard-coded impressions, though more interpretative, may introduce noise or inaccuracies that the model struggles to utilize effectively. The lack of improvement with speech impressions implies that allowing the model to learn directly from objective speech descriptions might be a more reliable approach.

Additionally, we compare the performance of numerical and categorical acoustic features using simple machine learning models, as detailed in Appendix A.2.

4.4.2 Can LLMs Leverage Speech Features Alone?

To further validate the effectiveness of simple speech features, we conduct experiments using only speech features in the prompt, as shown in Figure 3. The results in Table 3 demonstrate the potential of LLMs in leveraging these features for emotion classification, significantly outperforming

```
Now you are expert of sentiment and emotional
analysis using only audio features.

Target speech characteristics: very low volume
with very low variation, very high pitch with
very high variation, high speaking rate.

Please select the emotional label from <happy,
sad, neutral, angry, excited, frustrated> based
on the audio features. Respond with just one
label:
```

Figure 3: LLM Prompt Template for speech-feature-only Emotion Detection.

Method	Weighted F-1
blind estimation	16.67
numerical features (ML)	32.0
3-class categories (LLM)	27.602
5-class categories (LLM)	27.895
speech impression (LLM)	27.794

Table 3: Comparison of Weighted F-1 Scores for Different Classification Methods on IEMOCAP **using speech-feature only**

the blind estimation baseline of 16.67%. Specifically, the 5-class categorical representation with LLMs achieves a weighted F1 score of 27.895%, while the 3-class representation yields 27.602%, suggesting that decreasing the granularity of speech feature categorization does not significantly affect the LLM's ability to classify emotions.

As a reference point, using numerical feature representations with traditional machine learning (ML) models results in a weighted F1 score of 32%. The lower performance of the speech-only LLM approach compared to the ML approach highlights key differences in how these models process information. ML models, designed for discriminative tasks, can directly exploit structured, one-hot encoded categorical features. In contrast, LLMs interpret natural language descriptions, introducing a layer of abstraction and potential ambiguity. Despite this performance gap, LLMs' performance

Method Model	Text-only	Text + speech des	Text + speech imp
Claude Sonnet 3.5 (zero-shot)	58.99	**59.23** (+0.24)	59.11 (+0.12)
LLaMA-2-7b-base (LoRA)	70.111 ±0.36	**72.021** ±0.54 (+1.910)	71.542 ±1.12 (+1.431)
LLaMA-3-8b-base (LoRA)	70.316 ±0.95	71.744 ±0.79 (+1.428)	**71.910** ±0.62 (+1.594)
LLaMA-3-8b-instruct (LoRA)	70.474 ±0.94	**72.098** ±0.65 (+1.624)	71.132 ±0.36 (+0.658)
phi-3-13b-instruct (LoRA)	69.218 ±0.73	70.450 ±0.86 (+1.232)	**70.582** ±0.24 (+1.364)

Table 4: *SpeechCueLLM*performances (weighted F1 score in %) on IEMOCAP dataset across different models and input modalities. Results for LoRA fine-tuned models are reported as mean ± standard deviation over three runs. Bold indicates the best performance for each model. Dark green values in parentheses show the improvement over the Text-only method.

remains substantially above the blind estimation baseline, indicating that they can still extract meaningful emotional cues from the speech descriptions.

4.4.3 How Do LLMs Differ in Performance?

To evaluate the effectiveness of our approach across different language models, we conduct experiments using several popular open-sourced LLMs, including LLaMA-2, LLaMA-3, and Phi-3, along with a powerful closed-sourced model, Claude Sonnet 3.5. Table 4 presents the emotion recognition accuracy on the IEMOCAP dataset for these models under different input conditions: text-only, text with speech descriptions, and text with speech impressions.

Our results show consistent performance improvements across all models when either speech descriptions or impressions are integrated, underscoring the value of audio-derived features in emotion recognition tasks. In most cases, speech descriptions slightly outperform speech impressions, though the differences are often marginal. This suggests that more objective, feature-based descriptions may be more reliable than hard-coded speech impressions for emotion detection tasks.

Interestingly, increasing model size does not necessarily lead to better performance in this task. Despite their stronger performance on public benchmarks (Abdin et al., 2024; Dubey et al., 2024), larger models like LLaMA-3 and Phi-3 do not consistently outperform the smaller LLaMA-2 model when fine-tuned for emotion detection. For instance, Phi-3, the largest model tested (13B parameters), performs worse than both LLaMA-2 and LLaMA-3 across all input conditions. Although LLaMA-3-8b-instruct achieves the highest accuracy (72.098%) when provided with speech descriptions, the improvement over LLaMA-2-7b-base (72.021%) is minimal. The performance differences between the instruct and base versions of these models are insignificant for this task.

The substantial performance gap between the zero-shot Claude model and the fine-tuned models highlights the importance of task-specific fine-tuning for emotion recognition. However, even in a zero-shot setting, Claude benefits from the inclusion of speech features, further emphasizing the value of multimodal inputs in emotion detection.

5 Conclusions

In summary, our contributions are twofold: first, we introduce *SpeechCueLLM*, a novel approach that translates audio features into natural language descriptions to enable multimodal emotion analysis without architectural changes. This method significantly improves performance, especially with high-quality audio, and outperforms baseline models that rely on structural modifications.

Second, our key findings highlight the effectiveness and limitations of the method, emphasizing the importance of audio quality and the benefits of prompt-based techniques. Incorporating speech descriptions improves emotion recognition across LLMs, though the approach's success depends on audio quality, as shown by varying dataset results. Our experiments also demonstrate that prompt-based methods outperform direct integration of speech encoder features, leveraging LLMs' strengths in language processing.

Future work could prioritize on improving feature extraction techniques to better handle noisy, real-world audio data, ensuring robustness across diverse audio conditions. Moreover, advancing natural language representations of speech characteristics could further enhance both interpretability and accuracy in multimodal emotion recognition. Expanding this methodology to other domains and modalities also presents exciting opportunities for further research.

Limitations

While our approach to enhancing LLM-based emotion detection shows promise, it is important to recognize its limitations. A major constraint is the method's reliance on audio quality. As shown by the contrasting results between the IEMOCAP and MELD datasets, performance degrades significantly when the input audio is noisy or of low quality.

Another challenge is the nature of the training datasets. Our study primarily relies on acted (IEMOCAP) or scripted (MELD) emotional expressions, which may not fully capture the complexity and subtlety of emotions in natural, spontaneous speech. This dataset bias could limit the generalizability of our findings to real-world scenarios, where emotions are often more nuanced, mixed, or ambiguously expressed. The difference between performed emotions in our datasets and the diverse emotional expressions in everyday life may impact the system's effectiveness in real-world applications.

Finally, the computational resources required for this approach present a practical limitation. Fine-tuning large language models is computationally intensive, which may hinder the implementation of this method in resource-constrained environments. This requirement could restrict the accessibility and widespread adoption of our approach, especially in settings where high-performance computing resources are not readily available.

Addressing these limitations in future research will be critical for developing more robust, versatile, and scalable emotion recognition systems that can handle the complexities of real-world emotional expressions across diverse audio conditions and computational environments.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

M.Y. Alva, M. Nachamai, and Joy Paulose. 2015. A comprehensive survey on features and methods for speech emotion detection. In *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–6.

Mehdi Arjmand, Mohammad Javad Dousti, and Hadi Moradi. 2021. Teasel: A transformer-based speech-prefixed language model. *Preprint*, arXiv:2109.05522.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. *Preprint*, arXiv:2202.03555.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, Lu Gao, Yi Guo, Minglun Han, Ting Han, Wenchao Hu, Xinying Hu, Yuxiang Hu, Deyu Hua, Lu Huang, Mingkun Huang, Youjia Huang, Jishuo Jin, Fanliu Kong, Zongwei Lan, Tianyu Li, Xiaoyang Li, Zeyang Li, Zehua Lin, Rui Liu, Shouda Liu, Lu Lu, Yizhou Lu, Jingting Ma, Shengtao Ma, Yulin Pei, Chen Shen, Tian Tan, Xiaogang Tian, Ming Tu, Bo Wang, Hao Wang, Yuping Wang, Yuxuan Wang, Hanzhang Xia, Rui Xia, Shuangyi Xie, Hongmin Xu, Meng Yang, Bihong Zhang, Jun Zhang, Wanyi Zhang, Yang Zhang, Yawei Zhang, Yijie Zheng, and Ming Zou. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *Preprint*, arXiv:2407.04675.

Murali Karthick Baskar, Andrew Rosenberg, Bhuvana Ramabhadran, Neeraj Gaur, and Zhong Meng. 2024.

Speech prefix-tuning with rnnt loss for improving llm predictions. *Preprint*, arXiv:2406.14701.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Varshaa Dhanasekar, Yenugu Preethi, Vishali S, Praveen Joe I R, and Booma Poolan M. 2021. A chatbot to promote students mental health through emotion recognition. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1412–1416.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres,

Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2024. AudioChatLlama: Towards general-purpose speech abilities for LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532, Mexico City, Mexico. Association for Computational Linguistics.

Robert W Frick. 1985. Communicating emotion: The role of prosodic features. *Psychological bulletin*, 97(3):412.

Yumeng Fu. 2024. Ckerc : Joint large language models with commonsense knowledge for emotion recognition in conversation. *Preprint*, arXiv:2403.07260.

Lucas Goncalves and Carlos Busso. 2022. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, 13(4):2156–2170.

Taesik Gong, Josh Belanich, Krishna Somandepalli, Arsha Nagrani, Brian Eoff, and Brendan Jou. 2023. Lanser: Language-model supported speech emotion recognition. *arXiv preprint arXiv:2309.03978*.

Ziwei Gong, Muyin Yao, Xinyi Hu, Xiaoning Zhu, and Julia Hirschberg. 2024. A mapping on current classifying categories of emotions used in multimodal models for emotion recognition. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 19–28, St. Julians, Malta. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International journal of speech technology*, 15:99–117.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2024. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *Preprint*, arXiv:2309.11911.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023a. Large language models understand and can be enhanced by emotional stimuli. *Preprint*, arXiv:2307.11760.

Dongting Li, Chenchong Tang, and Han Liu. 2024. Audio-llm: Activating the capabilities of large language models to comprehend audio data. In *Advances in Neural Networks – ISNN 2024*, pages 133–142, Singapore. Springer Nature Singapore.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2024a. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26:776–788.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024b. An embarrassingly simple approach for llm with strong asr capacity. *Preprint*, arXiv:2402.08846.

Tao Meng, Fuchen Zhang, Yuntao Shou, Wei Ai, Nan Yin, and Keqin Li. 2024. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. *Preprint*, arXiv:2404.17862.

Sylvie Mozziconacci. 2002. Prosody and emotions. In *Speech Prosody 2002, International Conference*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *Preprint*, arXiv:1810.02508.

Jennifer Santoso, Kenkichi Ishizuka, and Taiichi Hashimoto. 2024. Large language model-based emotional speech annotation using context and acoustic feature for speech emotion recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11026–11030.

Rode Snehal Sudhakar and Manjare Chandraprabha Anil. 2015. Analysis of speech features for emotion detection: A review. In *2015 International Conference on Computing Communication Control and Automation*, pages 661–664.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. *Preprint*, arXiv:1906.00295.

Nguyen Van Thanh. 2023. Emotion recognition systems in retail: a detailed analysis of their role in enhancing customer interactions, driving sales, and predicting trends. *Journal of Computational Social Dynamics*, 8(3):1–9.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. 2024. Multimodal multi-loss fusion network for sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3588–3602, Mexico City, Mexico. Association for Computational Linguistics.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shixiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2023. Secap: Speech emotion captioning with large language model. *Preprint*, arXiv:2312.10381.

Jieying Xue, Minh Phuong Nguyen, Blake Matheny, and Le Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. *Preprint*, arXiv:2407.04279.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *Preprint*, arXiv:1707.07250.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024a. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. *Preprint*, arXiv:2402.13276.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2024b. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Preprint*, arXiv:2310.11374.

A Appendix

A.1 Contextual Speech Features

Method (IEMOCAP)	**Weighted F-1**
text-only	70.111 ±0.36
+ speech des	72.021 ±0.54
+ volume (all utterances)	71.226 ±0.25
+ volume (the last three)	71.487 ±0.59
+ pitch (all utterances)	71.28 ±0.47
+ pitch (the last three)	**72.595** ±0.26
+ all features (the last three)	71.461 ±0.76

Table 5: Comparison of Weighted F-1 Scores for Different Methods of Adding Speech-Feature Context on IEMOCAP. All experiments in the table use LoRA on the LLaMA-2-7b-base model.

In our exploration of incorporating speech features into context utterances, we found that selective application of these features yielded better results than a blanket approach. As shown in Table 5, adding speech descriptions to only the last three utterances in the context proved more effective than applying them to all utterances. This suggests that recent context plays a more significant role in emotion recognition. Notably, pitch features outperformed volume features, with pitch information from the last three utterances providing a 0.6% improvement over non-contextual speech descriptions, achieving a weighted F1 score of 72.595%.

Interestingly, combining all speech features for the context utterances did not enhance performance, likely due to information overload. These findings underscore the importance of selective feature inclusion and careful consideration of context in multimodal emotion recognition tasks, where focusing on recent and relevant speech features can improve results without overwhelming the model with excess information.

A.2 Evaluation of Speech Features Using Simple Machine Learning Models

To evaluate the effectiveness of our extracted speech features in capturing emotion-related information, we conducted experiments using various machine learning models on both numerical and categorical representations of these features. Tables 6 and 7 present the performance of Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) models on the IEMOCAP and MELD datasets.

Our analysis of the IEMOCAP dataset reveals that the five extracted speech features contain substantial emotion-related information. Both weighted F1 and macro F1 scores (approximately 0.32 and 0.30, respectively) are nearly double the random guess baseline of 0.167. This significant improvement over the baseline indicates that the speech features are effectively capturing emotional cues. Notably, converting numerical features into categorical features does not result in a significant loss of emotion-related information. Additionally, dividing the numerical features into five classes yields the best performance, with weighted F1 and macro F1 scores of 0.313 and 0.293, respectively, suggesting that a moderate level of discretization preserves clear emotional patterns in speech features.

The results from the MELD dataset highlight a crucial aspect of speech-based emotion recognition: the importance of audio quality. While weighted F1 scores (around 0.36) surpass the random guess baseline of 0.143, the macro F1 scores (approximately 0.15) are only marginally better than the baseline. This discrepancy is important: MELD is highly imbalanced, with neutral emotions comprising about 50% of the dataset. The high weighted F1 score primarily reflects the model's ability to predict the dominant neutral class, rather than its capacity to distinguish between other emotions from speech features. In contrast, the near-baseline macro F1 score, which treats all classes equally, reveals that the extracted speech features provide minimal informative value across diverse emotion categories.

This lack of substantial improvement in macro F1 scores indicates that the low quality and challenging nature of the audio in MELD are key factors in the poor performance. These findings underscore the critical importance of audio quality in obtaining reliable speech features for emotion recognition tasks. They also explain the modest gains observed when incorporating speech descriptions in our earlier language model experiments on MELD, in contrast with the significant improvements seen in IEMOCAP.

A.3 Additional Speech Features Processing Tables

A.4 Baseline Model Figure

Features	RF	SVM	MLP	**Average**
Random Guess	–	–	–	0.167(0.167)
Numerical	0.33(0.31)	0.33(0.31)	0.30(0.29)	**0.320**(**0.303**)
3-class	0.31(0.28)	0.32(0.30)	0.30(0.27)	0.310(0.283)
4-class	0.28(0.27)	0.32(0.29)	0.29(0.27)	0.297(0.277)
5-class	0.31(0.29)	0.33(0.30)	0.30(0.29)	0.313(0.293)
6-class	0.30(0.28)	0.31(0.29)	0.29(0.28)	0.300(0.283)

Table 6: Performance comparison of discriminant ML models using different feature representations on IEMOCAP using weighted F1(Macro F1) scores

Features	RF	SVM	MLP	**Average**
Random Guess	–	–	–	0.143(0.143)
Numerical	0.37(0.15)	0.34(0.12)	0.38(0.18)	0.363(0.150)
3-class	0.37(0.15)	0.36(0.13)	0.37(0.16)	**0.367**(0.147)
4-class	0.36(0.15)	0.35(0.13)	0.36(0.14)	0.357(0.140)
5-class	0.36(0.16)	0.35(0.13)	0.36(0.15)	0.357(0.147)
6-class	0.36(0.17)	0.35(0.13)	0.35(0.17)	0.353(**0.157**)

Table 7: Performance comparison of discriminant ML models using different feature representations on MELD using weighted F1(Macro F1) scores

Number of Classes	**Quantile Splits**
3	[0.25, 0.75]
4	[0.25, 0.5, 0.75]
5	[0.1, 0.25, 0.75, 0.9]
6	[0.1, 0.25, 0.5, 0.75, 0.9]

Table 8: Quantile splits for different numbers of classes

Feature	**Level**	**Impression**
Pitch	High/Very High	Uses a higher pitch
	Low/Very Low	Uses a lower pitch
	Medium	Has a moderate pitch
Pitch Variation	High/Very High	With noticeable variation, suggesting expressiveness
	Low/Very Low	That remains steady, potentially indicating calmness or seriousness
	Medium	With typical variation
Volume	High/Very High	Speaking loudly, which might indicate excitement, confidence, or urgency
	Low/Very Low	Speaking softly, possibly suggesting calmness, shyness, or caution
	Medium	Using a moderate volume
Volume Variation	High/Very High	With significant volume changes
	Low/Very Low	With little volume variation
	Medium	With normal volume variation
Speech Rate	High/Very High	Talking quickly, which could indicate excitement, urgency, or nervousness
	Low/Very Low	Talking slowly, possibly suggesting thoughtfulness, hesitation, or calmness
	Medium	Speaking at a moderate pace

Table 9: Feature-Impression Mapping

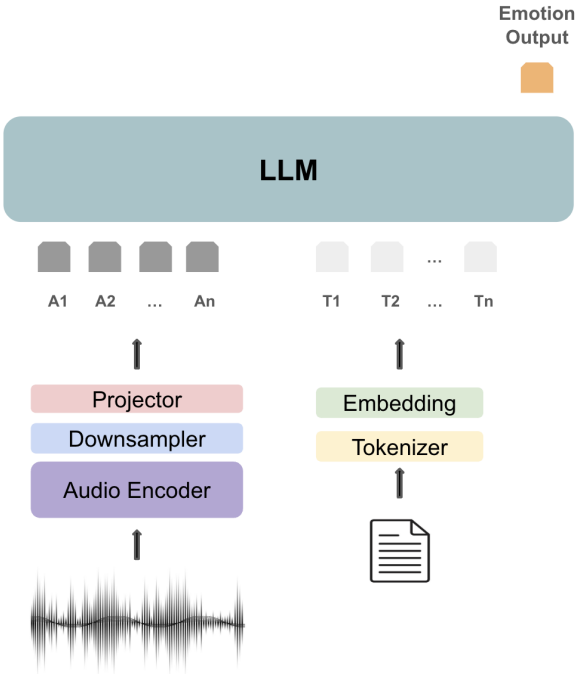

Figure 4: The baseline model structure involves projecting speech encoder features directly into the LLM embedding space.