Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

AA-SLLM: An Acoustically Augmented Speech Large Language Model for Speech Emotion Recognition

Jialong Mai[1], *Xiaofen Xing*[1,*], *Weidong Chen*[3], *Yuanbo Fang*[1], *Xiangmin Xu*[1,2]

[1]School of Electronic and Information Engineering, South China University of Technology, China
[2]School of Future Technology, South China University of Technology, China
[3]Systems Engineering & Engineering Management, The Chinese University of Hong Kong, China

202320111090@mail.scut.edu.cn, xfxing@scut.edu.cn, wdchen@se.cuhk.edu.hk,
eeybfang@mail.scut.edu.cn, xmxu@scut.edu.cn

Abstract

Recently, the rapid advancements in Speech Large Language Models (SpeechLLMs) have greatly accelerated progress in the Speech Emotion Recognition (SER) field. However, SpeechLLMs rely on powerful semantic encoders and acoustically irrelevant pre-training data, granting limited attention to acoustic information, which is closely related to the emotion in speech. In this paper, we leverage acoustic properties correlated with emotions to automatically generate acoustic descriptions. These descriptions are combined with the semantic representations as inputs to the LLM, enhancing emotion recognition capabilities. Accordingly, we propose AA-SLLM, an acoustically augmented SpeechLLM adopting instruction fine-tuning via Low-Rank Adaptation (LoRA). Experimental results indicate that AA-SLLM effectively alleviates the class imbalance problem while improving overall performance. Furthermore, AA-SLLM achieves state-of-the-art results on IEMOCAP, MELD, and LSSED datasets.

Index Terms: speech emotion recognition, speech large language model

1. Introduction

Human emotion plays a critical role in communication. Speech Emotion Recognition (SER) is an essential tool for informing intelligent systems about users' feelings [1, 2] and is widely used in practical applications, such as intelligent robots, automated call centers, and distance education [3, 4].

Researchers are exploring ways to expand Large Language Models (LLMs) into speech processing applications [5, 6, 7, 8, 9]. This integration has achieved performance comparable to or exceeding traditional expert models across various tasks, such as SER, Speech Language Understanding (SLU), Automatic Speech Recognition (ASR), and so on [10, 8, 9, 11]. Building on this, recent work has targeted SER within SpeechLLMs [12], leveraging the robust semantic understanding capabilities of LLMs to achieve significant performance gains. However, despite these advancements, a notable performance gap between these models and human-level emotion recognition persists, indicating the need for further exploration in leveraging SpeechLLMs for SER.

Numerous studies have confirmed that many acoustic features, such as pitch, intensity, speech rate, and articulation rate, are highly correlated with emotion [13, 14, 15]. For example, changes in signal intensity correspond to shifts in emotional states [16, 17, 18]. More intense emotions such as anger or excitement are usually associated with higher signal intensity, while more depressive emotions such as sadness or depression have lower signal energy. Recent research on SER has also demonstrated that incorporating explicit acoustic information effectively improves the recognition accuracy [19, 20, 21].

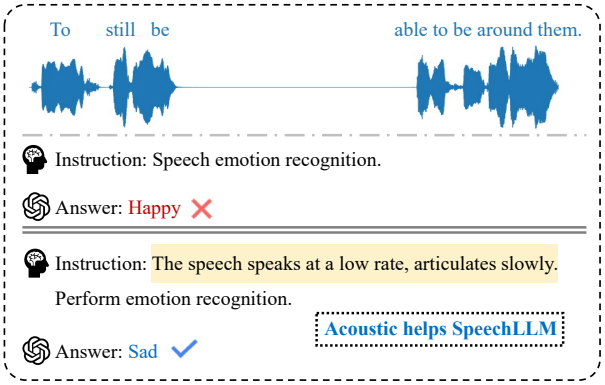

Figure 1: *A case study in LSSED dataset showing the benefits of utilizing acoustic information for SER task.*

Nevertheless, despite the importance of acoustic features, current SpeechLLMs [10, 22, 23, 24, 9, 7] solely use semantic audio encoders, such as the Whisper encoder [25], pretrained on ASR tasks, to strengthen semantic comprehension. The promising acoustic features related to emotion have not been sufficiently considered. In addition, some studies [26, 10] incorporate audio-caption pairs during the pretraining stage, the captions are acoustically independent, often presenting broad, macroscopic depictions of sound scenes, events, and speaker gender (e.g., "A female voice communicates over a static walkie-talkie" [27]). These typical captions lack detailed information about low-level acoustic features associated with emotion and are suboptimal for SER task.

To address the above issues, we propose a method to explicitly leverage low-level acoustic information in SpeechLLMs to enhance SER. Specifically, we extract emotional acoustic variations from each utterance and use them to construct the descriptions automatically, which are then combined with the audio semantic representations. The combined features, containing both acoustic and semantic information, serve as input to LLM to assist in emotion recognition. We introduce AA-SLLM, an acoustically augmented SpeechLLM that jointly models acoustic and semantic information by employing Low-Rank Adaptation (LoRA) [28] for instruction fine-tuning. AA-SLLM is trained on the large-scale emotion dataset LSSED [29]. The experimental results show that AA-SLLM effectively alleviates the class imbalance problem while improving recognition accuracy.

*Corresponding author.

10.21437/Interspeech.2025-1915

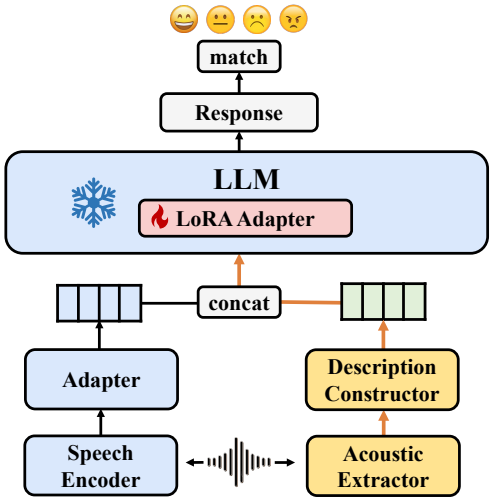

Figure 2: *Overview structure of the proposed AA-SLLM.*

In summary, the contributions of this paper are as follows:

- We propose an automatic paradigm to generate acoustic descriptions in an efficient manner.
- We introduce AA-SLLM, an acoustically augmented SpeechLLM that jointly models acoustic and semantic information by adopting LoRA for instruction fine-tuning, enhancing its capability for emotion recognition.
- The experimental results show that AA-SLLM effectively alleviates the class imbalance problem while improving recognition accuracy. Furthermore, AA-SLLM achieves state-of-the-art results on IEMOCAP, MELD, and LSSED datasets.

2. Methodology

2.1. SpeechLLM for SER

Basically, a traditional SER model processes audio signals $\mathbf{X} = [x_1, \ldots, x_T]$ as its input and infers the output one-hot code $\mathbf{R} = [0, \ldots, 1, \ldots, 0]$, where the position of the '1' indicates the predicted emotion category.

A standard SpeechLLM architecture consists of three core components: a speech encoder, an adapter module, and a foundational text-based LLM. Initially, audio signals $\mathbf{X}$ are input into the speech encoder, generating $\mathbf{H} = \text{SpeechEncoder}(\mathbf{X})$. Then, the adapter module maps $\mathbf{H}$ into $\mathbf{H}'$ via $\mathbf{H}' = \text{Projection}(\mathbf{H})$, aligning it with the dimensionality of the LLM's textual embeddings, the adapter is a fully-connected layer. $\mathbf{H}'$ is subsequently fed into the LLM along with various textual prompts, which are designed to direct the LLM toward specific downstream tasks. For instance, using the prompt "speech emotion recognition" with a special token $\langle\text{SER}\rangle$ is effective for the SER task. The LLM then generates task-specific outputs, denoted as z_i, in an autoregressive manner.

$$z_i \leftarrow \text{LLM}(H', \langle\text{SER}\rangle, z_{<i-1}). \tag{1}$$

To reduce SpeechLLM training costs, researchers incorporate pretrained modules for most components and employ Low-Rank Adaptation (LoRA) to optimize training efficiency. The speech encoder is initialized with the Whisper-large-v3, while foundational models such as Qwen serve as the LLM's initialization. The training objective is defined as a multi-class cross-entropy loss applied to each predicted token.

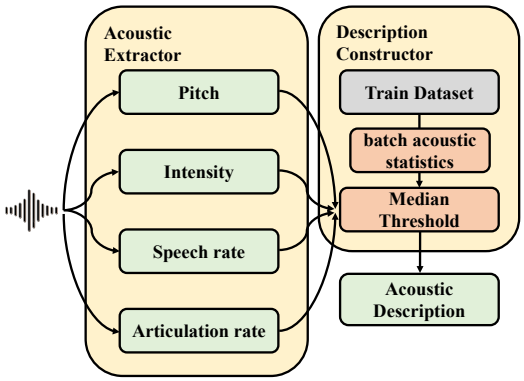

Figure 3: *The details of Acoustic Extractor and Description Constructor.*

2.2. AA-SLLM

As shown in Figure 2, the proposed AA-SLLM consists of five main components: LLM, Speech Encoder, Adapter, Acoustic Extractor, and Description Constructor.

2.2.1. Acoustic Extractor (AE)

Acoustic properties play a crucial role in SER, as they provide essential information about the speaker's emotional state. Inspired by [19], we focus on four key acoustic properties: pitch, intensity, speech rate, and articulation rate. These features capture prosodic and temporal information that is closely related to emotional expressions.

Pitch, reflecting the perceived frequency of a speaker's voice, serves as a vital indicator of emotional states [14].

To compute pitch, the audio signal X is first analyzed using the Short-Time Fourier Transform (STFT), and then the pitch values are extracted via the piptrack algorithm [30].

The average pitch of the whole utterance is computed as:

$$\text{Pitch} = \frac{1}{N}\sum_{k=1}^{N}\text{piptrack}(X, sr) \tag{2}$$

where sr is the sampling rate, N is the number of non-zero pitch values.

Intensity, which measures the loudness of the speech signal, is a crucial feature for differentiating between emotional states [14].

It is computed as the root mean square (RMS) energy of the audio signal:

$$\text{Intensity} = 10 \cdot \log_{10}\left(\frac{\sum_{n=1}^{L} X[n]^2}{L}\right) \tag{3}$$

where $X[n]$ is the audio signal, and L is the length of the sample.

Speech Rate is the number of spoken syllables per unit of time, capturing the rhythm of speech. This is a crucial temporal feature for SER [13]. To compute speech rate, the audio signal is partitioned into syllables via energy-based segmentation:

$$\text{Syllables} = \text{split}(X, \text{top\_db} = 20) \tag{4}$$

where Syllables represent the set of segmented syllables, this means that a signal is only considered a valid syllable or word

if its energy exceeds the background noise by 20 dB. The speech rate is then computed as:

$$\text{Speech Rate} = \frac{\text{num(Syllables)}}{\text{Duration(X)}} \tag{5}$$

Articulation Rate is a finer temporal indicator, representing the speed of spoken syllables during active phonation (excluding pauses) [13]. Phonation time is computed by summing the durations of active syllables:

$$\text{Phonation Time} = \sum_{k=1}^{num(Syllables)} \frac{\text{End}[k] - \text{Start}[k]}{\text{sr}} \tag{6}$$

where End$[k]$ and Start$[k]$ are the boundaries of the k-th syllable. The articulation rate is then given by:

$$\text{Articulation Rate} = \frac{\text{num(Syllables)}}{\text{Phonation Time}} \tag{7}$$

2.2.2. Description Constructor (DC)

To enable SpeechLLM to interpret the acoustic properties extracted via the Acoustic Extractor, we introduce the Description Constructor, a method that summarizes acoustic features using dataset-derived statistical thresholds to ensure adaptability across domains and mitigate domain variance.

Specifically, as illustrated in Figure 3, the thresholds for each acoustic property are determined from the training set using the median value of the respective feature across all samples. Let $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ represents the training dataset, where each sample d_i contains four acoustic properties: pitch, intensity, speech rate, and articulation rate. For a given acoustic property $f(d)$, its threshold is calculated as:

$$\text{Threshold}_f = \text{Median}\big(\{f(d_i) \mid d_i \in \mathcal{D}\}\big), \tag{8}$$

where Median$(\cdot)$ denotes the statistical median.

For each test sample, the Description Constructor evaluates extracted acoustic features against predefined thresholds, generating categorical labels ("Low" for below-threshold values or "High" for above-threshold values) to characterize each feature.

We use standard next token prediction mechanism to train our AA-SLLM and apply Low-Rank Adaptation (LoRA) technique to reduce the training burden.

3. Experiments

To train AA-SLLM, we utilized a large-scale emotional dataset LSSED [29]. This extensive dataset provides a rich source of acoustic variability and diverse emotional expressions, making it ideal for training AA-SLLM. AA-SLLM effectively addresses the class imbalance issue, demonstrating strong recognition capabilities for minority emotion classes. In the following sections, we outline the principles guiding our dataset selection, the setup of our experiments, and conduct comparative and ablation studies to thoroughly evaluate our method's effectiveness.

3.1. Datasets

To fully leverage the potential of an existing SpeechLLM, our approach builds Qwen2-Audio-7B [10] pretrained on extensive pairs of speech-transcription and audio-caption data. Prior to our instruction fine-tuning, Qwen2-Audio-7B has not undergone task-specific tuning for emotion recognition. To ensure minimal exposure of the pre-trained model to emotion labels, we carefully select datasets excluded from the pretraining process of this model.

IEMOCAP [31] is used following the same way as in previous studies [32, 33, 34]. We merge excitement into the happiness category and select 5,531 utterances from happy, angry, sad and neutral classes. The experiments are conducted using the leave-one-session-out cross-validation strategy. We use IEMOCAP because the paper [10] explicitly states that it was not used in the training process of Qwen2-Audio.

MELD [35] comprises 13,708 utterances across 7 emotion classes. It is officially divided into training, validation, and testing sets. We use the validation set for hyperparameters tuning, and present the scores on the testing sets. We use MELD because the Qwen2-Audio-Instruct model employed it to evaluate emotion recognition capabilities, ensuring that label leakage during pretraining was impossible.

LSSED [29] is, to the best of our knowledge, one of the largest speech emotion datasets available, which has data collected from 820 subjects and contains 147,025 samples. Consistent with [34], we use LSSED with four emotion categories: angry, neutral, happy, and sad. LSSED is a proprietary dataset we constructed, containing data from 820 subjects and 147,025 samples. Access to this dataset requires authorization, and its use is strictly regulated. As we have not signed an EULA with Qwen's development organization, the dataset's security remains fully ensured.

3.2. Experiment Setup

We performed instruction-tuning on Qwen2-Audio-7B for five epochs, leveraging Whisper-Large-v3 [25] as the audio encoder. Low-Rank Adaptation (LoRA) was applied with a rank of 8 and alpha of 32 to all linear layers in the LLM and Adapter, optimizing parameter efficiency while minimizing computational costs. All non-LoRA parameters remained frozen, resulting in 4.19M trainable parameters. Training employed the AdamW optimizer with a learning rate of 1e-4 and a cosine annealing scheduler (5% warmup ratio). Batch sizes were set to 1 for IEMOCAP and MELD, and 32 for LSSED. Experiments were executed on an NVIDIA A100 GPU with 80GB VRAM.

3.3. Experimental Results and Analysis

3.3.1. Comparison with Some Known Systems

1. **Pre-trained models**

   Pre-trained model representations form the foundation for expert models and LLM-based approaches. We compare WavLM [36], Whisper [25], and emotion2vec [37] as pre-trained baselines. To ensure parameter comparability with our model, we froze these pre-trained models and appended a single 1024-dimensional Transformer encoder layer, resulting in 5.26M trainable parameters, marginally exceeding our model's 4.19M.

2. **Expert models**

   DropFormer and EMER [32, 38] serve as the latest expert model for SER tasks, based on the pre-trained model. Expert models build upon pre-trained models by enhancing downstream model architectures or employing other techniques to improve the specialization of pre-trained models for emotion recognition tasks.

3. **LLM-based models**

   SELM [12] is the latest LLM-based model specifically designed for SER tasks. Moreover, SALMONN [8] is a LLM-

based model integrated with multiple downstream tasks, as a comparative example.

AA-SLLM outperforms on all three datasets, as shown in Table 1. Pre-trained models perform worse than expert models. This is because expert models typically tailor their downstream network structures to the specific characteristics of emotion recognition tasks. LLM-based methods demonstrate significant performance advantages, likely attributable to their large-scale semantic alignment training and extensive parameter capacity enabling generalized understanding.

Table 1: *Comparison with state-of-the-art systems on IEMOCAP (UA), MELD (WF1), and LSSED (UA).*

Method	Year	IEMOCAP	MELD	LSSED
WavLM [36]	2022	69.47	46.52	42.70
SELM [12]	2024	73.09	-	-
emotion2vec [37]	2024	73.20	48.70	44.69
Whisper [25]	2022	73.54	51.45	46.60
DropFormer [32]	2024	76.60	49.25	44.56
EMER [38]	2024	77.16	-	-
Salmonn [8]	2023	-	53.34	-
Ours	2025	**85.50**	**55.23**	**50.20**

3.3.2. Ablation Study

We analyzed the ablation results of the IEMOCAP, MELD, and LSSED datasets. Among these, LSSED exhibits an imbalanced distribution of emotional categories, with Neutral as the predominant class to reflect real-world scenarios. In the training samples, instances labeled as Neutral comprise 56.79% of the total. Therefore, we adopt UA as the evaluation metric.

Table 2: *Ablation results of the AA-SLLM's core components on the IEMOCAP, MELD and LSSED.*

IEMOCAP	
Model	UA
AA-SLLM	**85.50**
AA-SLLM w/o AE & DC	84.35
AA-SLLM w/o LLM	73.54
MELD	
Model	WF1
AA-SLLM	**55.23**
AA-SLLM w/o AE & DC	53.60
AA-SLLM w/o LLM	51.45
LSSED	
Model	UA
AA-SLLM	**50.20**
AA-SLLM w/o AE & DC	47.77
AA-SLLM w/o LLM	46.60

In the Table 2, AE stands for Acoustic Extractor, and DC stands for Description Constructor. We implemented a model without using acoustic augmentation strategies, denoted as "AA-SLLM w/o AE & DC". When employing the acoustic augmentation scheme, UA shows an increase, attributed to the model's enhanced ability to recognize minority categories.

The "w/o LLM" setup excludes the large language model, using audio encoder features for emotion recognition. Since the audio encoder architecture is identical in both models, the improved performance of AA-SLLM demonstrates the benefits of decomposing the SER task into an audio encoder followed by language model reweighting.

3.3.3. Fairness in Emotion Recognition: Addressing Class Imbalance

We focused on the samples misclassified by the model, as they highlight the model's limitations. Neutral samples make up the largest proportion in the model's training data, prompting us to investigate whether the model exhibits a bias toward misclassifying minority class samples as neutral. As illustrated in Table 3, among samples where AA-SLLM was correct but the model without acoustic features erred, 820 samples (70.21%) were wrongly labeled as Neutral by the latter. In contrast, when the model without acoustic features classified samples correctly but AA-SLLM misclassified them, only 327 samples (31.72%) were wrongly labeled as Neutral by AA-SLLM.

Table 3: *The left side of the table shows the samples misclassified as neutral by the model without the acoustic strategy, while the right side shows those misclassified by our model with the acoustic strategy.*

w/o AE & DC×		**AA-SLLM**×	
True label	Count	True label	Count
Happy	621	Happy	177
Sad	94	Sad	131
Angry	105	Angry	19
Sum	820	Sum	327

This indicates that when training samples are imbalanced, SpeechLLMs are prone to misclassifying instances from minority classes as belonging to the majority class (e.g., Neutral). However, integrating the acoustic strategy, this issue can be effectively mitigated. This is because the low-level acoustic features of samples from different categories inherently exhibit significant distinctions, remaining unaffected by the imbalance in sample sizes. Therefore, the model can learn to enhance the recognition accuracy of minority classes from an acoustic perspective, thereby alleviating the impact of class imbalance.

This helps explain why AA-SLLM demonstrates a greater improvement on the LSSED dataset. LSSED has a highly imbalanced class distribution, whereas IEMOCAP and MELD also exhibit imbalanced class distributions (with the largest class being Neutral, accounting for 25% and 33.2%, respectively). However, their degree of imbalance is much lower than that of LSSED.

4. Conclusions

In this paper, we propose AA-SLLM, an acoustically augmented SpeechLLM that models acoustic and semantic information jointly by adopting LoRA for instruction fine-tuning. Traditional SpeechLLMs often overlook acoustic information, which is crucial for emotion recognition, due to their reliance on acoustically irrelevant pre-training data and semantic-focused encoders. By integrating acoustic features such as pitch, intensity, speech rate, and articulation rate, our method mitigates the above limitation, generating detailed acoustic descriptions that complement the semantic audio inputs to the language models. The experimental results demonstrate that AA-SLLM not only effectively alleviates the class imbalance problem but also significantly improves performance. Our approach achieves state-of-the-art results on IEMOCAP, MELD, and LSSED.

5. Acknowledgements

The work is supported in part by Guangdong Basic and Applied Basic Research Foundation 2025A1515011203; in part by Guangdong Provincial Key Laboratory of Human Digital Twin 2022B1212010004.

6. References

[1] J. J. Gross, H. Uusberg, and A. Uusberg, "Mental illness and well-being: an affect regulation perspective," *World Psychiatry*, vol. 18, no. 2, pp. 130–139, 2019.

[2] J. J. Gross and R. F. Muñoz, "Emotion regulation and mental health." *Clinical psychology: Science and practice*, vol. 2, no. 2, p. 151, 1995.

[3] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6675–6679.

[4] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[5] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.

[6] S. Hu, L. Zhou, S. Liu, S. Chen, L. Meng, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran *et al.*, "Wavllm: Towards robust and adaptive speech large language model," *arXiv preprint arXiv:2404.00656*, 2024.

[7] Y. Shu, S. Dong, and C. et al., "Llasm: Large language and speech model," *arXiv preprint arXiv:2308.15930*, 2023.

[8] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[9] Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma, W. Wang, S. Zheng *et al.*, "Lauragpt: Listen, attend, understand, and regenerate audio with gpt," *arXiv preprint arXiv:2310.04673*, 2023.

[10] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[11] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue."

[12] H. Bukhari, S. Deshmukh, H. Dhamyal, B. Raj, and R. Singh, "Selm: Enhancing speech emotion recognition for out-of-domain scenarios," 2024.

[13] R. W. Frick, "Communicating emotion: The role of prosodic features." *Psychological bulletin*, vol. 97, no. 3, p. 412, 1985.

[14] K. R. Scherer, "Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences." 1972.

[15] A. Pavlenko, "Emotions and multilingualism," 2005.

[16] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals." in *Interspeech*. Citeseer, 2003, pp. 125–128.

[17] K. Venkataramanan and H. R. Rajamohan, "Emotion recognition from speech," *arXiv preprint arXiv:1912.10458*, 2019.

[18] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[19] H. Dhamyal, B. Elizalde, S. Deshmukh, H. Wang, B. Raj, and R. Singh, "Prompting audios using acoustic properties for emotion representation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 936–11 940.

[20] Z. Wu, Z. Gong, L. Ai, P. Shi, K. Donbekci, and J. Hirschberg, "Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances," *arXiv preprint arXiv:2407.21315*, 2024.

[21] J. Santoso, K. Ishizuka, and T. Hashimoto, "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)*, 2024, pp. 11 026–11 030.

[22] Y. Li, X. Wang, S. Cao, Y. Zhang, L. Ma, and L. Xie, "A transcription prompt-based efficient audio large language model for robust speech recognition," *arXiv preprint arXiv:2408.09491*, 2024.

[23] X. Gong, A. Lv, Z. Wang, and Y. Qian, "Contextual biasing speech recognition in speech-enhanced large language model," *Interspeech*, 2024.

[24] Z. Xie and C. Wu, "Mini-omni: Language models can hear, talk while thinking in streaming," *arXiv preprint arXiv:2408.16725*, 2024.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[26] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," *arXiv preprint arXiv:2305.10790*, 2023.

[27] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[29] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "Lssed: a large-scale dataset and benchmark for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 641–645.

[30] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.

[31] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[32] J. Mai, X. Xing, W. Chen, and X. Xu, "Dropformer: A dynamic noise-dropping transformer for speech emotion recognition," in *Proc. Interspeech 2024*, 2024, pp. 2645–2649.

[33] Z. Li, X. Xing, Y. Fang, W. Zhang, H. Fan, and X. Xu, "Multi-scale temporal transformer for speech emotion recognition," *arXiv preprint arXiv:2410.00390*, 2024.

[34] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[35] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[36] S. Chen, C. Wang, and Z. e. a. Chen, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1505–1518, Oct. 2022.

[37] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 15 747–15 760.

[38] J. Kyung, S. Heo, and J.-H. Chang, "Enhancing multimodal emotion recognition through asr error compensation and llm fine-tuning," in *Proc. Interspeech 2024*, 2024, pp. 4683–4687.