



Chain-of-Thought Distillation with Fine-Grained Acoustic Cues for Speech Emotion Recognition

Jialong Mai¹, Xiaofen Xing^{1,*}, Yangbiao Li¹, Xiangmin Xu^{1,2}

¹School of Electronic and Information Engineering, South China University of Technology, China

²School of Future Technology, South China University of Technology, China

202320111090@mail.scut.edu.cn, xfxing@scut.edu.cn, 202420164005@mail.scut.edu.cn, xmxu@scut.edu.cn

Abstract

Recent advances in large language models (LLMs) have demonstrated strong reasoning abilities through chain-of-thought (CoT) prompting, yet their application in speech emotion recognition (SER) remains underexplored. Moreover, current SER models lack explainability based on emotion-related acoustic features. We propose AECoTD, a method that transfers reasoning abilities from a large LLM to a domain-specific SER LLM by leveraging fine-grained emotional acoustic features and text transcripts. It uses LoRA to distill the reasoning chain and an emotion-focused loss to preserve correct emotional attention, thereby enhancing the model's explainability. Ablation experiments highlight the impact of fine-grained acoustic information, emotional CoT reasoning, and emotion-focused loss. Without using pre-trained representations, our method achieves state-of-the-art performance both in-domain and out-of-domain, demonstrating strong generalization ability.

Index Terms: speech emotion recognition, chain-of-thought reasoning

1. Introduction

Human emotion plays a crucial role in communication. Speech emotion recognition (SER) is an essential tool for enabling intelligent systems to understand users' emotional states [1, 2]. It has become widely applied in various domains, including intelligent robots, automated call centers, and distance education [3, 4].

Chain-of-Thought advancements in large language models (LLMs), such as DeepSeek-R1 [5], have revolutionized NLP tasks by providing transparent reasoning processes. However, their potential in speech emotion recognition (SER) remains underexplored. Current SER systems [6, 7] based on one-hot labels often fail to model contextual affective dependencies, suggesting an opportunity to integrate LLMs' interpretable reasoning chains. To address this gap, developing methods that explicitly map speech signals to structured reasoning steps could improve both model performance and interpretability, as transparent emotion inference logic may reveal overlooked acoustic-semantic correlations that are critical for accurate recognition.

Numerous studies have established that acoustic features, such as pitch and intensity, are closely linked to emotional states [8, 9, 10]. Research has shown that utterance-level acoustic information is crucial [11]. However, fine-grained segment-level acoustic cues within an utterance are equally important. As illustrated in Figure 1, the speech transcript "I am going to the store" appears neutral in text. On the other hand, the fine-grained acoustic features for pitch and intensity progressively

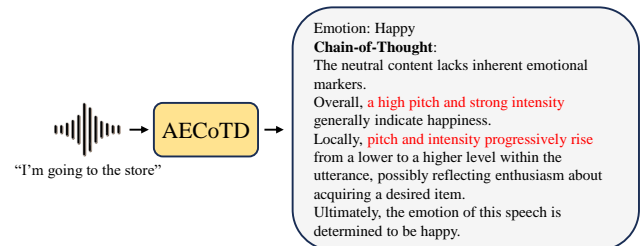


Figure 1: AECoTD leverages fine-grained acoustic features to derive the emotional chain-of-thought from DeepSeek-R1, thereby significantly enhancing the explainability of the model.

rise from a lower to a higher level, possibly indicating excitement about purchasing a desired item. Nevertheless, aligning implicit temporal acoustic information with text remains a challenge. Large language models (LLMs) with chain-of-thought reasoning capabilities offer a promising solution, as they can model the sequential dependencies between acoustic variations and textual semantics, providing interpretable reasoning steps for emotion inference.

Many studies apply the interpretability of large language models to SER [12, 13, 14, 15]. For instance, SECap [12] uses LLaMA [16] to generate brief emotional descriptions of audio, but these descriptions lack detailed reasoning. Additionally, its descriptions are based on the utterance level, without considering the intrinsic acoustic variations relevant to emotion. AffectGPT [13] leverages multimodal inputs to infer emotions. However, it employs ImageBind [17] as a pre-processing module for the audio modality, which indirectly aligns audio with text. This process prevents AffectGPT from incorporating acoustic information in a way that LLMs can effectively understand, particularly fine-grained acoustic variations within an utterance.

To address the above challenges, we propose Acoustic Emotional Chain-of-Thought Distillation (AECoTD), an innovative two-stage framework.

(1) AECoTD generates interpretable emotion-driven reasoning paths by integrating both coarse-grained and fine-grained acoustic features with text transcripts. Coarse-grained features capture the overall acoustic characteristics of an utterance, informing the LLM of how the utterance's global acoustic properties compare to those of other utterances (e.g., whether it is generally louder, higher-pitched). Fine-grained features, on the other hand, represent the acoustic characteristics of individual segments within the utterance, allowing the LLM to understand how each segment's acoustic properties compare to others within the same utterance (e.g., detecting rising pitch, sudden intensity shifts). This enables the model to capture in-

*Corresponding author.

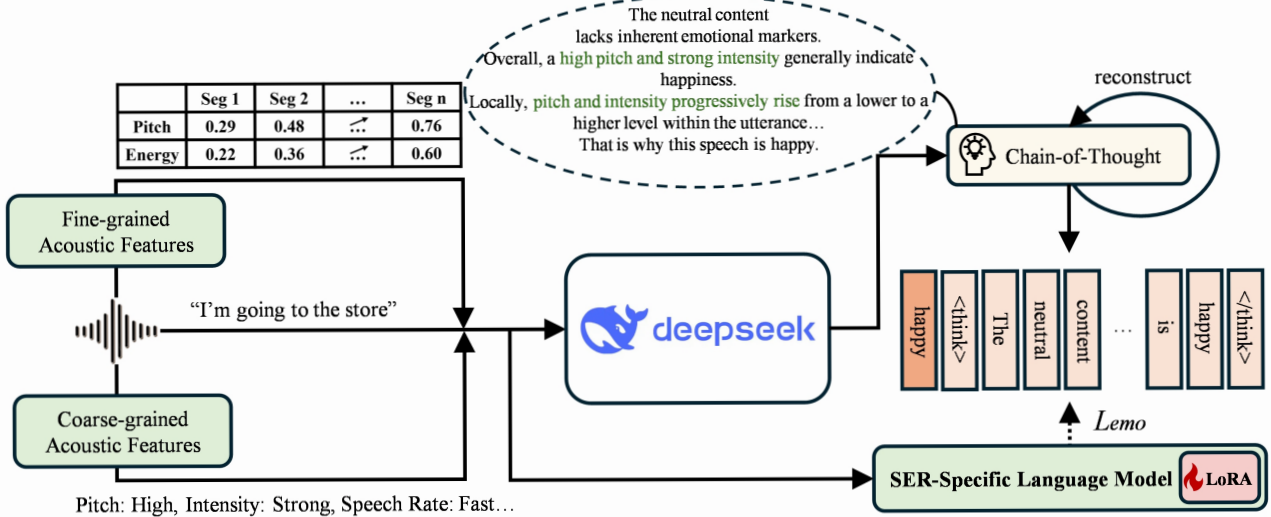


Figure 2: The AECoTD framework. L_{emo} represents the emotion-focused loss we designed, where the dark-colored token (happy) indicates the parts that the model should pay more attention to, while the lighter-colored tokens represent the parts that are of secondary importance to the model.

ternal acoustic dynamics. Leveraging DeepSeek-R1’s reasoning ability, AECoTD jointly models textual semantics and dynamic acoustic information, constructing an emotional chain-of-thought based on acoustic cues.

(2) To efficiently transfer the distilled reasoning chain, we fine-tune Qwen2.5-7B [18] using Low-Rank Adaptation (LoRA) [19]. Additionally, we introduce a specially designed emotion-focus loss, which encourages the model to concentrate on reasoning towards the correct emotion, mitigating the risk of attention being diluted by excessively long reasoning chains.

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

- We are the first to apply chain-of-thought reasoning to acoustic-based emotional inference, thereby significantly enhancing the explainability of the model.
- We introduce Acoustic Emotional Chain-of-Thought Distillation (AECoTD), which adopts LoRA to distill the reasoning chain into a domain-specific language model. Additionally, we incorporate an emotion-focused loss to minimize the dilution of attention on the correct emotion caused by excessively long reasoning chains.
- Experimental results highlight the impact of fine-grained acoustic information, emotional CoT reasoning, and the emotion-focused loss in ensuring the attention remains on the correct emotion. Without using pre-trained representations, our method achieves state-of-the-art performance both in-domain and out-of-domain, demonstrating strong generalization ability.

2. Methodology

The AECoTD framework operates in two phases, as shown in Figure 2. First, it combines multi-level acoustic features with text transcripts, leveraging DeepSeek-R1 to construct emotion-driven chain-of-thought. Second, it distills the reasoning chain into Qwen2.5-7B through LoRA fine-tuning and employs an emotion-focus loss to enhance reasoning towards the correct emotion while preventing attention dilution from overly long reasoning chains.

2.1. Coarse-grained Acoustic Features

Coarse-grained features capture the overall acoustic characteristics of an utterance, informing the LLM how its global acoustic properties compare to those of other utterances. We focus on four key acoustic features: pitch, intensity, speech rate, and articulation rate. These properties capture prosodic and temporal information that closely correlates with emotional expressions.

To compute pitch, we first perform an STFT on the audio signal and then extract pitch using the piptrack [20] algorithm. The average pitch is determined from the non-zero values.

Intensity is measured by calculating the root mean square (RMS) energy of the audio and converting it to a decibel scale.

For speech rate, the audio is segmented into syllables using an energy-based method with a 20 dB threshold above background noise. The speech rate is then obtained by dividing the number of syllables by the audio duration.

Articulation rate reflects the speed of syllable production during active speech (excluding pauses) and is calculated by dividing the number of syllables by the total phonation time.

We propose a dataset-driven thresholding strategy to capture how each utterance’s acoustic properties compare to others. For each acoustic feature (pitch, intensity, speech rate, and articulation rate), we calculate a threshold based on the training set by computing the median value across all samples. Let $\mathcal{D} = \{d_1, d_2, \dots, d_N\}$ 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:

$$\text{Threshold}_f = \text{Median}(\{f(d_i) \mid d_i \in \mathcal{D}\}), \quad (1)$$

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

We then compare each sample’s acoustic features to these thresholds and assign coarse-grained labels (e.g., high/low for pitch, strong/weak for intensity, and fast/slow for speech rate or articulation rate).

2.2. Fine-grained Acoustic Features

Global metrics for pitch and intensity can miss important local variations over time. Fine-grained features, however, capture these subtle changes and complement the broader, coarse-grained measures.

These fine-grained features describe the acoustic properties of individual segments within an utterance, allowing the LLM to discern how each segment differs from others. Since speech transcripts and acoustic features are not strictly aligned, using frame-level features directly is impractical due to their length and complexity. To address this, we propose the following extraction framework:

Firstly, divide the audio signal into non-overlapping temporal segments. For each segment, compute the mean pitch and intensity (denoted as F_i^{mean} for the i -th segment) as described previously. This approach preserves local prosodic details that would be lost in whole-signal averaging.

Secondly, to simplify the LLM’s task, we normalize the features within each segment by determining their upper and lower bounds. A sliding window (0.5s window size, 0.25s step size) is applied across the entire signal to establish feature bounds, which are calculated as follows:

$$F_{\max} = \max_{1 \leq k \leq K} (F^{\text{window}}(k)) \quad (2)$$

where K denotes the total number of sliding windows. F_{\max} represents the upper bound of the acoustic feature values for the current utterance, while $F^{\text{window}}(k)$ denotes the mean acoustic feature value within the k -th sliding window. The calculation of the lower bound of the feature values F_{\min} is similar.

By calculating per-window extremal values, we obtain stable global maxima and minima that are less affected by brief irregularities like pulse spikes. For each segment i with mean acoustic feature F_i , we apply min-max normalization as:

$$\hat{F}_i = \frac{F_i^{\text{mean}} - F_{\min}}{F_{\max} - F_{\min}} \quad (3)$$

where \hat{F}_i denotes the normalized feature value of the i -th segment and F_i^{mean} representing the mean feature value of the i -th segment.

This operation linearly projects each feature dimension into the unit interval $[0, 1]$, with silent segments (identified by intensity thresholds) automatically mapped to zero values.

For each sample, the fine-grained feature values are presented in a list, showing the normalized mean values for different time segments, which can be understood by the LLM. We only compute the fine-grained feature values for pitch and intensity, as changes in these two features provide more effective information.

2.3. AECoTD

We leverage DeepSeek-R1’s reasoning ability to jointly model textual semantics and dynamic acoustic information, constructing an emotional chain-of-thought based on acoustic cues. In our approach, DeepSeek-R1 receives both coarse-grained and fine-grained acoustic features along with the corresponding audio transcripts. Using these inputs, we craft a set of instructions to elucidate emotions. DeepSeek-R1 then generates a chain-of-thought that captures the dynamic acoustic-based reasoning underlying emotion recognition. This process can be formalized as follows:

$$C = R_{\text{DeepSeek-R1}} \left(T, \{F_j\}_{j=1}^4, \left\{ \{\hat{F}_{j,i}\}_{i=1}^n \right\}_{j=1}^2 \right) \quad (4)$$

where C represents the output chain-of-thought generated by DeepSeek-R1, $R_{\text{DeepSeek-R1}}(\cdot)$ denotes the reasoning function implemented by DeepSeek-R1, T is the text transcript of the audio sample, $\{F_j\}_{j=1}^4$ represents the set of four coarse-grained acoustic features extracted from the utterance, and $\left\{ \{\hat{F}_{j,i}\}_{i=1}^n \right\}_{j=1}^2$ denotes the collection of fine-grained acoustic feature lists for the two selected features (typically pitch and intensity), with each $\{\hat{F}_{j,i}\}_{i=1}^n$ being the list of normalized feature values computed over n temporal segments.

We restructured DeepSeek-R1’s output to better serve as a label for Qwen2.5-7B in the format: `emotion label<think>chain-of-thought</think>`.

However, during distillation, we found that using a conventional cross-entropy loss—which distributes attention evenly across all tokens—poses a problem when the chain-of-thought is much longer than the emotion label. This can cause the model to focus more on the detailed explanation rather than on the correct emotion.

To overcome this, we designed an emotion-focused loss that adjusts token weights. This approach preserves the acoustic-based reasoning while emphasizing correct emotion deduction. We assign a weight of 1 to the emotion label and 0.01 to all other tokens. The final loss function is formulated as follows:

$$\mathcal{L} = - \sum_t w_t \cdot y_t \log \hat{y}_t \quad (5)$$

where y_t and \hat{y}_t represent the ground truth and predicted probabilities for token t , respectively. The token weight w_t is defined as:

$$w_t = \begin{cases} 1, & \text{if } t \text{ belongs to the emotion label} \\ 0.01, & \text{otherwise} \end{cases} \quad (6)$$

3. Experiments

3.1. Datasets

IEMOCAP [21] is used as in previous studies [6, 7]. We merge excitement with happiness and select 5,531 utterances from the happy, angry, sad, and neutral classes, using a leave-one-session-out cross-validation strategy.

MELD [22] consists of 13,708 utterances across 7 emotion classes. Here, we use MELD to evaluate out-of-domain performance and select samples from happy, angry, sad, and neutral categories following Emobox [23].

3.2. Experiment Setup

Using the acoustic emotion chain-of-thought from DeepSeek-R1, we fine-tuned Qwen2.5-7B with LoRA for 10 epochs. LoRA was applied to all linear layers in the LLM with a rank of 8 and alpha of 32, optimizing parameter efficiency while reducing computational cost. Training employed the AdamW optimizer with a 1e-4 learning rate and a cosine annealing scheduler (5% warmup), with a batch size of 1. Experiments were conducted on an NVIDIA A100 GPU with 80GB VRAM.

3.3. Experimental Results and Analysis

3.3.1. Ablation Study

This section presents the ablation study results on the IEMOCAP dataset, evaluating the contribution of each component in AECoTD. As shown in Table 1, systematically removing key

elements reveals their impact on performance. All audio segments containing fine-grained acoustic strategies have a segment count of $n = 7$.

Model	UA
AECoTD	72.21
AECoTD w/o CoT	70.84
AECoTD w/o CoT & F	68.78
AECoTD w/o CoT & F & C	65.39
AECoTD w/o Emoloss	64.13

Table 1: Ablation results of the AECoTD’s core components on the IEMOCAP. *F* denotes fine-grained acoustic features, while *C* represents coarse-grained acoustic features.

The Chain-of-Thought (CoT) mechanism improves performance by 1.37% UA, demonstrating its role in explicit emotion reasoning. Fine-grained acoustic features (F) contribute a 2.06% UA gain by capturing temporal variations in pitch and intensity, while coarse-grained features (C) add 3.39% UA by establishing a global acoustic context. Notably, removing the emotion-focused loss (EmoLoss) leads to the most significant drop (8.08% UA), highlighting its importance in balancing attention between emotion labels and lengthy reasoning chains. Without emotion-focused loss, the model struggles to prioritize emotion deduction over explanatory text generation, resulting in substantial performance degradation.

3.3.2. Comparison Study

In this section, we compare the in-domain and out-of-domain performance of our model with state-of-the-art SER models. To the best of our knowledge, existing studies on interpretable emotion recognition have not conducted experiments on IEMOCAP, making direct comparisons impossible. Table 2 presents the in-domain performance on the IEMOCAP dataset.

Method	Year	Text	Audio	UA
AECoTD	2025	✓		72.21
Vesper [24]	2024		✓	70.80
MMRBN [25]	2024		✓	70.20
WavLM-Large [26]	2022		✓	69.47
Hubert-Large [27]	2021		✓	67.42
LoRA-MER [28]	2024	✓		60.94

Table 2: In-Domain Performance Comparison on IEMOCAP.

Our experiments show that AECoTD, which leverages both text and a small set of precisely extracted acoustic cues (converted to text), performs on par with state-of-the-art SER models that rely on pre-trained acoustic features. In contrast, LoRA-MER—when limited to text—lags behind AECoTD by 11.27% UA, underscoring the benefits of our acoustic emotion reasoning chain. By integrating acoustic reasoning with emotion recognition, AECoTD not only matches state-of-the-art performance but also offers superior interpretability.

Out-of-domain tests are essential for assessing SER’s real-world robustness against varying emotion definitions and distributions. Table 3 shows AECoTD’s out-of-domain performance (trained on IEMOCAP and tested on MELD) under the Emobox setting, where MELD is restricted to the four emotion classes defined in IEMOCAP. Due to the lack of a unified benchmark and diverse experimental setups, we compare only with publicly available pre-trained models.

Method	WA
AECoTD	54.59
Whisper-Large-V3 [29]	51.42
Hubert-Large [27]	44.69
Data2vec 2.0 [30]	41.75
WavLM-Large [26]	39.06

Table 3: Out-of-Domain Performance Comparison on MELD.

AECoTD significantly outperforms all methods in out-of-domain experiments, highlighting its exceptional generalization ability and real-world applicability. This advantage likely stems from the compact, domain-specific language model acquiring reasoning skills from DeepSeek-R1, combined with fine-grained acoustic cues to perform temporally aware emotional inference. Moreover, semantic and acoustic features are far less affected by domain differences compared to pre-trained representations, further reinforcing AECoTD’s strong generalization capability.

3.3.3. Hyperparameter Study

The ablation study confirms the effectiveness of fine-grained acoustic features, making it crucial to analyze the impact of n , which determines their granularity. As shown in Figure 4, we conducted experiments on IEMOCAP without the chain-of-thought mechanism to examine how LLMs interpret acoustic information at different segmentation levels.

n	1	3	5	7	9
UA	68.78	69.59	70.00	70.84	70.51

Table 4: Impact of fine-grained segmentation level on UA performance in IEMOCAP

When $n = 1$, the model receives only utterance-level acoustic information, leading to a significant performance drop due to the lack of detailed acoustic cues. As segmentation becomes finer, the model captures more subtle acoustic variations linked to emotional expression, improving performance. However, excessively fine segmentation may not provide additional benefits, as acoustic features and text are not strictly aligned, and complex inputs may hinder the LLM’s ability to generate coherent reasoning chains. AECoTD achieves the best balance with $n = 7$, effectively capturing acoustic trends while maintaining semantic coherence, leading to the most informative chain-of-thought.

4. Conclusions

We propose Acoustic Emotional Chain-of-Thought Distillation (AECoTD), a method that integrates fine-grained acoustic features with text to transfer reasoning capabilities from DeepSeek-R1 into a smaller, domain-specific SER LLM. AECoTD employs LoRA for efficient reasoning distillation and introduces an emotion-focused loss to ensure attention remains on the correct emotion rather than overly long reasoning chains. Ablation studies confirm the effectiveness of fine-grained acoustic features, emotional CoT reasoning, and emotion-focused loss. Without relying on pre-trained representations, AECoTD achieves state-of-the-art performance both in-domain and out-of-domain, demonstrating strong generalization.

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 and R. F. Muñoz, "Emotion regulation and mental health." *Clinical psychology: Science and practice*, vol. 2, no. 2, p. 151, 1995.
- [2] 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.
- [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. Y. H. Z. J. S. R. Z. DeepSeek-AI, Daya Guo *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: <https://arxiv.org/abs/2501.12948>
- [6] 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.
- [7] 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.
- [8] R. W. Frick, "Communicating emotion: The role of prosodic features." *Psychological bulletin*, vol. 97, no. 3, p. 412, 1985.
- [9] K. R. Scherer, "Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences." 1972.
- [10] A. Pavlenko, "Emotions and multilingualism," 2005.
- [11] H. Dharmyal, 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.
- [12] Y. Xu, H. Chen, J. Yu, Q. Huang, Z. Wu, S.-X. Zhang, G. Li, Y. Luo, and R. Gu, "Secap: Speech emotion captioning with large language model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 323–19 331.
- [13] Z. Lian, H. Sun, L. Sun, J. Yi, B. Liu, and J. Tao, "Affect-gpt: Dataset and framework for explainable multimodal emotion recognition," *arXiv preprint arXiv:2407.07653*, 2024.
- [14] Y. Xu, Y. Zhou, Y. Cai, J. Xie, R. Ye, and Z. Wu, "Multimodal emotion captioning using large language model with prompt engineering," in *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, 2024, pp. 104–109.
- [15] Z. Lian, L. Sun, M. Xu, H. Sun, K. Xu, Z. Wen, S. Chen, B. Liu, and J. Tao, "Explainable multimodal emotion reasoning," *arXiv preprint arXiv:2306.15401*, 2023.
- [16] K. S. P. A. A. Hugo Touvron, Louis Martin *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: <https://arxiv.org/abs/2307.09288>
- [17] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.
- [18] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," 2024. [Online]. Available: <https://arxiv.org/abs/2406.07162>
- [24] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2024.
- [25] X. Chen, "Mmrbn: Rule-based network for multimodal emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8200–8204.
- [26] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [27] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021. [Online]. Available: <https://arxiv.org/abs/2106.07447>
- [28] Y. Cai, Z. Wu, J. Jia, and H. Meng, "Lora-mer: Low-rank adaptation of pre-trained speech models for multimodal emotion recognition using mutual information," in *Proc. Interspeech 2024*, 2024, pp. 4658–4662.
- [29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [30] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.