MSF-SER: ENRICHING ACOUSTIC MODELING WITH MULTI-GRANULARITY SEMANTICS FOR SPEECH EMOTION RECOGNITION

Haoxun Li[1] *Yuqing Sun*[1] *Hanlei Shi*[1] *Yu Liu*[1] *Leyuan Qu*[⋆,1] *Taihao Li*[⋆,1]

[1] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences

ABSTRACT

Continuous dimensional speech emotion recognition captures affective variation along valence, arousal, and dominance, providing finer-grained representations than categorical approaches. Yet most multimodal methods rely solely on global transcripts, leading to two limitations: (1) all words are treated equally, overlooking that emphasis on different parts of a sentence can shift emotional meaning; (2) only surface lexical content is represented, lacking higher-level interpretive cues. To overcome these issues, we propose MSF-SER (Multi-granularity Semantic Fusion for Speech Emotion Recognition), which augments acoustic features with three complementary levels of textual semantics—Local Emphasized Semantics (LES), Global Semantics (GS), and Extended Semantics (ES). These are integrated via an intra-modal gated fusion and a cross-modal FiLM-modulated lightweight Mixture-of-Experts (FM-MOE). Experiments on MSP-Podcast and IEMOCAP show that MSF-SER consistently improves dimensional prediction, demonstrating the effectiveness of enriched semantic fusion for SER.

Index Terms— Speech Emotion Recognition, Dimensional Emotion Regression, Multi-Granularity Semantics, Prosodic Emphasis

1. INTRODUCTION

Speech Emotion Recognition (SER) has attracted increasing attention for its applications in human–computer interaction, mental health, and multimedia retrieval. Traditional approaches often rely on discrete emotion categories, which cannot fully capture the richness of human emotions. Recently, emotion dimensions have become popular and widely adopted, since they are more capable of capturing subtle variations and representing complex emotional states[1].

Early dimensional SER systems largely relied on handcrafted acoustic and prosodic features combined with conventional machine learning models[2]. Recent advances include the TF-Mamba architecture proposed by Zhao et al.[3] for capturing temporal and frequency patterns, and efficient fine-tuning strategies for large SER models introduced by Aneesha et al[4].

Despite progress in speech-only approaches, semantic representations derived from acoustic signals remain limited, as they are easily affected by noise, speaker variability, and pronunciation differences[5]. Text provides a more stable source of semantics, and pre-trained text models are trained on much larger corpora than speech models, yielding richer and more reliable representations[6]. This motivates the incorporation of textual semantics to complement acoustic features. However, most existing multimodal studies still rely solely on global transcripts, which introduces two major issues.

(1) Global transcripts treat all words equally, overlooking the fact that not every word contributes to emotion and that emphasis on different parts of a sentence can shift its affective meaning. For example, in the sentence "I really didn't mean that", placing emphasis on "really" conveys a sense of sincerity or insistence, while stressing "didn't" instead conveys denial or even defensiveness. Such differences can drastically alter the perceived emotion. This motivates the introduction of locally emphasized semantics, which are derived from the prosodically most prominent segments of speech. These emphasized regions often convey the speaker's key semantic intent and exhibit stronger correlations with emotional expression.

(2) Global transcripts only capture the explicit lexical content while lacking higher-level interpretive information, including but not limited to explanatory cues, situational factors, and paralinguistic attributes. Such information is essential for emotion recognition, since affective meaning is shaped not only by lexical content but also by inferred intent, communicative context, and speaker-specific characteristics. To address this gap, we incorporate prior-knowledge semantics, where external audio–language models provide extended and multi-dimensional interpretations of the audio, enhancing textual semantics with information inferred from audio. Along with local emphasized semantics and global contextual semantics, these prior-knowledge representations constitute our multi-granular semantic design.

To fully exploit enriched semantics, MSF-SER employs an intra-modal gated fusion to adaptively integrate different semantic levels, and a FiLM-modulated lightweight Mixture-of-Experts (FM-MOE) for cross-modal integration. Together, these components allow semantic cues to effectively guide acoustic representation learning, leading to consistent improvements in dimensional emotion recognition.

The main contributions of our work are summarized as follows:

- For speech emotion recognition, we supplement acoustic modeling with richer semantic information by extracting features at three complementary granularities.
- We employ FM-MOE for cross-modal fusion, while an intra-modal gating mechanism adaptively integrates global and local semantics to enhance acoustic feature learning.
- MSF-SER achieves top-performing results across MSP-Podcast and IEMOCAP, thereby demonstrating the effectiveness and generalizability of our method.

⋆ Corresponding authors.

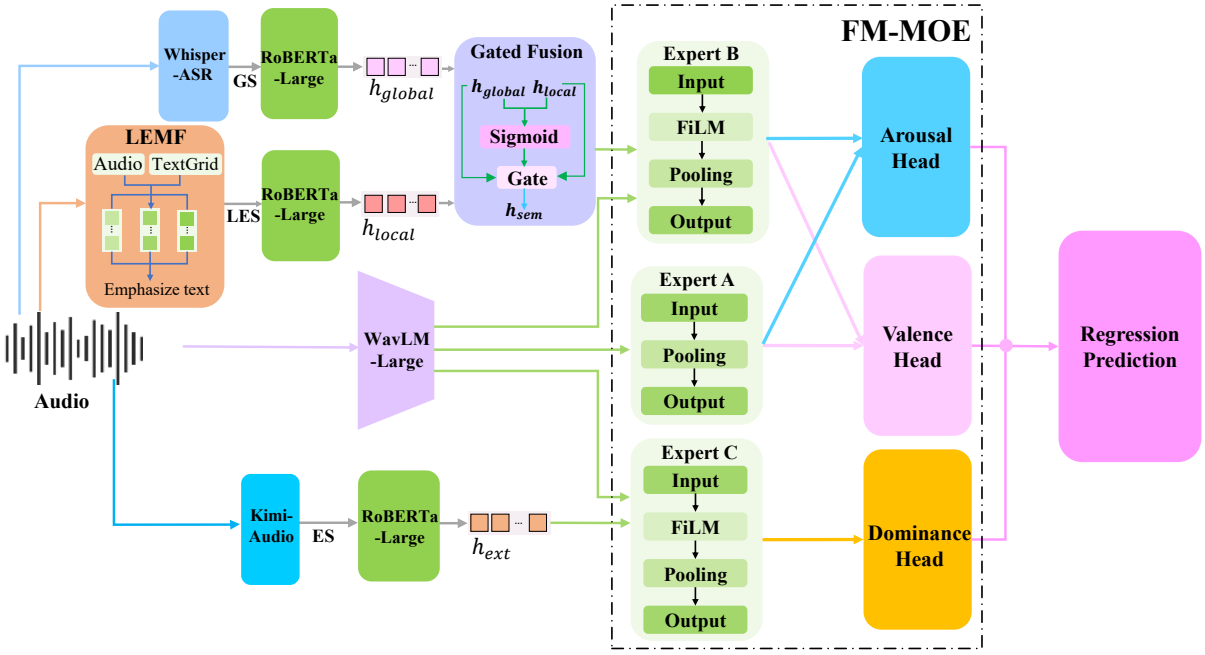

Fig. 1: Our proposed framework adopts acoustic features as the backbone, while three levels of semantic features are incrementally incorporated as auxiliary signals. An intra-modal gating mechanism fuses global and local semantics, and cross-modal integration between acoustic and semantic representations is achieved through FM-MOE.

2. METHOD

2.1. Overview

Our baseline employs a fine-tuned WavLM-large architecture, where raw audio is first encoded into frame-level acoustic representations by the WavLM encoder. The outputs are then aggregated using attentive statistics pooling to compute weighted mean and standard deviation, followed by a fully connected layer that predicts valence, arousal, and dominance.

On this foundation, we introduce multi-granularity semantic features: LES, GS, and ES, as illustrated in Fig. 1. These features interact with acoustic features through a two-stage fusion framework. For intra-modal fusion, we design a gated mechanism to adaptively integrate LES and GS, enhancing model sensitivity and stability. For inter-modal fusion, we propose FM-MOE. Specifically, Feature-wise Linear Modulation (FiLM) applies dimension-wise scaling and shifting to acoustic representations conditioned on semantic features, ensuring fine-grained cross-modal interaction. Within FM-MOE, FiLM-modulated experts specialize in different emotional dimensions, and their outputs are adaptively combined through different routing weights. This design enables semantic cues to effectively guide acoustic learning while capturing dimension-specific dependencies.

2.2. Multi-Granularity Semantic Feature Extraction

Local Emphasized Semantics (LES). Prior work indicates that prosodic prominence mainly arises from pitch, energy, and duration[7]. However, in noisy conditions, emphasis detection models trained on clean data, such as EmphaClass[8], often degrade. To make emphasis modeling more robust, we design a **Local Emphasis Modeling Framework (LEMF)**, which identifies prosodically prominent words and transforms them into LES. LES highlight the most salient parts of speech that often carry the speaker's key intent and emotional salience. The specific implementation of LEMF is as follows:

Given a sentence $u = \{w_1, \ldots, w_N\}$ with audio signal $x(t)$, where t denotes time, and its MFA-based TextGrid alignment[9], we extract three prosodic features for each word w:

- **Pitch** ($f_{\text{pitch}}(w)$): log-F0 via the PyWorld Harvest algorithm;
- **Energy** ($f_{\text{energy}}(w)$): L2 norm of Short-Time Fourier Transform (STFT) using a 20ms window;
- **Duration** ($f_{\text{duration}}(w)$): mean phoneme duration in w.

To ensure comparability of features within a sentence, we compute sentence-level statistics (mean μ_i, standard deviation σ_i) for pitch $f_{\text{pitch}}(w)$ and energy $f_{\text{energy}}(w)$. Each word-level feature is then normalized using Z-score standardization, as shown in Equation 1.

$$z_i(w) = \frac{f_i(w) - \mu_i}{\sigma_i}, \quad i \in \{\text{pitch}, \text{energy}, \text{duration}\} \tag{1}$$

where $f_i(w)$ denotes the raw feature of w in the i-th prosodic dimension. Specifically, pitch corresponds to the maximum frame-level F0, energy to the mean frame energy, and duration to average phoneme length.

Subsequently, we perform a weighted fusion of the three prosodic features to obtain the emphasis score s(w) for each word w as shown in Equation 2:

$$s(w) = \alpha \cdot z_{\text{pitch}}(w) + \beta \cdot z_{\text{energy}}(w) + \gamma \cdot z_{\text{duration}}(w) \tag{2}$$

where $(\alpha, \beta, \gamma) = (1.0, 1.2, 0.8)$. The word with the highest score, along with its two adjacent words, forms the emphasis segment. This segment is then encoded by RoBERTa-Large[10] to derive LES features.

Global Semantics (GS). We derive sentence-level semantics by transcribing speech with Whisper-ASR[11] and encoding the text using RoBERTa-Large.

Extended Semantics (ES). Leveraging recent advances in large-scale audio understanding, we employ Kimi-Audio[12] to generate six categories of extended speech and emotion-related information, as shown in Table 1. These outputs are concatenated into descriptive text and encoded with RoBERTa-Large, yielding enriched semantic representations. For example, an extended semantic description may state: "This is a male speaker, expressing frustrated (categorized as Angry), in a confrontation or argument in a public setting. The speaker is expressing anger due to a perceived lack of understanding or compliance from the other party. The speech is characterized by rising intonation, assertive tone, and a sense of urgency."

Table 1: Extended semantics information categories generated by Kimi-Audio.

Category	Description
Free emotion label	Open-domain emotion prediction
Constrained emotion label	Mapping to dataset labels
Emotion explanation	Reason for emotion prediction
Scenario	Potential situational context
Paralinguistics	Paralinguistic information
Gender	Gender information

2.3. Intra-modal Gated Fusion

Locally emphasized semantics capture salient emotional cues but lack holistic coverage, whereas global semantics provide comprehensive context but may introduce redundancy and noise. To integrate their complementary strengths, we employ a gated fusion mechanism, as shown in Equation 3.

$$h_{\text{sem}} = g \cdot h_{\text{local}} + (1 - g) \cdot h_{\text{global}} \tag{3}$$

$$g = \sigma(W[h_{\text{local}}; h_{\text{global}}] + b) \tag{4}$$

where h_{sem} is the fused representation, h_{local} and h_{global} denote local and global semantics, respectively, g is the learned gating weight, σ the Sigmoid activation, and W, b are trainable parameters.

2.4. FM-MOE

In dimensional emotion recognition, acoustic features remain the primary source of information, while text semantics provide complementary cues. To facilitate fine-grained cross-modal interaction, we employ a Feature-wise Linear Modulation (FiLM) layer[13], which applies dimension-wise scaling and shifting to acoustic representations conditioned on semantic features, as shown in Equation 5. This design prioritizes speech as the dominant modality while leveraging semantics as auxiliary guidance, thereby mitigating the impact of semantic noise and improving fusion effectiveness.

$$\tilde{h}_{\text{audio}} = \gamma \odot h_{\text{audio}} + \beta \tag{5}$$

$$\gamma, \beta = \text{MLP}(h_{\text{sem}}) \tag{6}$$

where $\tilde{h}_{\text{audio}}$ denotes the semantically modulated acoustic feature, h_{audio} represents the input acoustic feature, γ and β are the scaling and bias parameters learned through the Multi-Layer Perceptron (MLP), and h_{sem} is the semantic feature vector.

To capture the distinct cross-modal dependencies of different emotional attributes, we propose a lightweight Mixture-of-Experts[14] module following the shared WavLM backbone. It consists of three experts: an *acoustic-only* expert (Expert A), a *textual semantic* expert (Expert B) incorporating fused local and global semantics, and an *extended semantic* expert (Expert C) that integrates additional prior knowledge.

Subsequently, for each emotional dimension $d \in \{\text{V}, \text{A}, \text{D}\}$, the output is a weighted combination of the experts, as shown in Equation 7.

$$\hat{y}^{(d)} = \sum_{k=1}^{3} \pi_k^{(d)} \cdot f_k(h) \tag{7}$$

where f_k denotes the k-th expert and $\pi_k^{(d)}$ represents the learnable routing weight. The routing weights allow different emotional dimensions to adaptively emphasize relevant experts (e.g., Valence and Arousal typically rely more on Experts A and B, while Dominance emphasizes Experts C).

3. EXPERIMENTS

3.1. Experimental Setup

We conducted experiments on MSP-Podcast v1.12 and IEMOCAP. MSP-Podcast v1.12 is a large-scale corpus of spontaneous podcast speech with 84,260 training and 31,961 development utterances, annotated with ten categorical labels and dimensional ratings (arousal, valence, dominance, 1–7 scale), and officially split into training, development, and three test sets. IEMOCAP contains 10,039 utterances from ten actors across five sessions of scripted and spontaneous dialogues, annotated with categorical emotions and dimensional ratings on a 1–5 scale.

We train with AdamW (1×10^{-5} learning rate), batch size 32, and gradient accumulation of 4. Each emotional dimension is predicted by an independent two-layer regression head with dropout 0.5 and layer normalization. For features, we use WavLM-Large as the acoustic encoder (hidden size 1024), freezing the CNN front-end and fine-tuning the Transformer layers, and RoBERTa-Large (1024 hidden size) for textual semantics. Training is performed on 8 RTX 4090 GPUs.

3.2. Performance on Emotional Attribute Prediction

Our evaluation results on the MSP-Podcast dataset are reported on the development set, as shown in Table 2. Model performance is measured using the Concordance Correlation Coefficient (CCC), which jointly accounts for both correlation and mean squared difference between predictions and ground truth, making it the standard metric for dimensional emotion recognition.

We adopt a baseline architecture built on WavLM-Large as the speech encoder, followed by an attentive statistics pooling layer and fully connected layers to perform regression prediction. The results demonstrate that incorporating semantic information at three different granularities consistently improves recognition performance. Specifically, introducing GS and LES yields greater gains for arousal and valence prediction, while ES mainly enhances dominance regression. Regarding inter-modal fusion strategies, both concatenation and attention degrade performance. We attribute this to the fact that, in speech-dominant tasks, these methods allow noisy textual semantics to interfere with the intrinsic structure of acoustic representations. In contrast, FiLM-based fusion effectively mitigates such noise and leads to more robust predictions. Moreover, applying intra-modal fusion on GS and LES further improves results compared to using either individually, with the gated approach outperforming the attention-based method. Finally, when all three levels of semantic information are jointly introduced and integrated with FM-MOE, our model achieves the best overall performance.

We further evaluated our model on the IEMOCAP dataset to validate its effectiveness. We adopt five-fold cross-validation, using four sessions for training and the remaining session for testing

Table 2: Ablation results on the MSP-Podcast dataset.

Semantic			Intra-Modal Fusion		Inter-Modal Fusion				Evaluation Metrics			
GS	LES	ES	Attention	Gated	Concat	Attention	FiLM	FM-MOE	CCC_V	CCC_A	CCC_D	CCC_{avg}
									0.725	0.660	0.592	0.659
✓					✓				0.710	0.658	0.591	0.653
✓						✓			0.697	0.655	0.587	0.646
✓							✓		0.741	0.668	0.608	0.672
	✓						✓		0.739	0.665	0.610	0.671
		✓					✓		0.728	0.652	0.630	0.670
✓	✓		✓				✓		0.745	0.670	0.606	0.677
✓	✓			✓			✓		0.756	0.675	0.612	0.681
✓	✓	✓		✓			✓		0.750	0.670	0.622	0.681
ML-Adapt[15]									0.647	0.680	0.616	0.648
LoRA-SER[16]									0.701	0.682	0.616	0.666
DEER[17]									0.688	0.696	0.612	0.665
PCM-le-noNorm[18]									0.658	0.675	0.626	0.653
SERNC Top-Performing Model[19]									0.758	0.683	0.615	0.685
✓	✓	✓		✓				✓	**0.759**	**0.685**	**0.631**	**0.692**

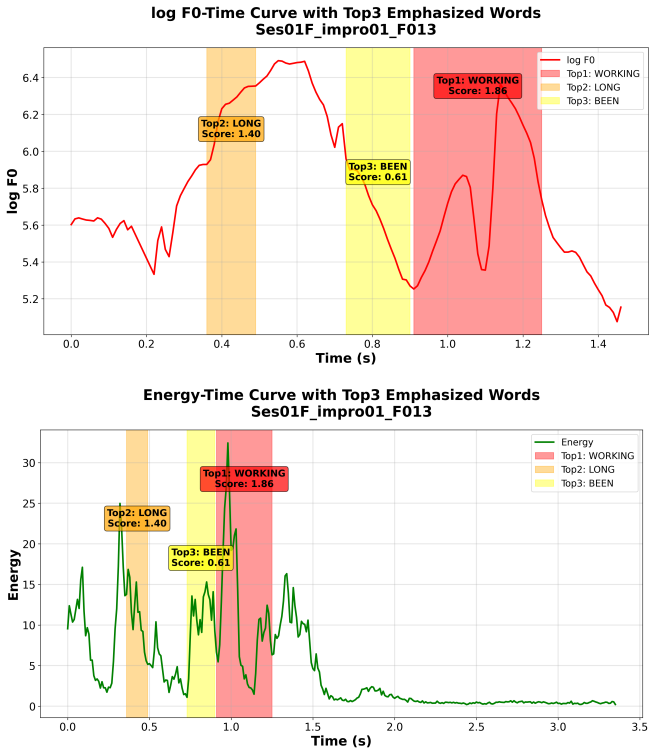

Fig. 2: Illustration of emphasis detection by LEMF. The top figure shows the log-F0 curve over time with the top-3 emphasized words highlighted, while the bottom figure shows the corresponding energy-time curve.

Table 3: Comparison of different models on the IEMOCAP dataset.

Model	**CCC_V**	**CCC_A**	**CCC_D**	**CCC_{avg}**
Baseline [20]	0.552	0.678	0.583	0.604
KNN-VC[21]	0.568	0.656	0.485	0.570
WavLM-LR[22]	0.625	0.675	0.599	0.633
DEER [17]	0.625	0.711	0.548	0.628
PCM-le-noNorm[18]	0.630	**0.717**	0.555	0.634
MSF-SER	**0.632**	0.680	**0.601**	**0.638**

in each fold, ensuring strict speaker independence. As shown in Table 3, we compared our approach with multiple models. MSF-SER achieves the highest CCC scores on valence, dominance and the overall average, demonstrating its advantage in capturing subtle emotional cues across dimensions.

3.3. LEMF

LEMF is designed to improve emphasis detection in spontaneous speech, where noise and speaking style variations often degrade performance. By leveraging prosodic features such as pitch, energy, and duration, LEMF can more accurately capture emphasized regions, yielding higher accuracy and robustness. As illustrated in Fig. 2 ,this feature-based approach not only enhances adaptability in spontaneous conditions but also provides reliable local semantic cues for downstream emotion recognition tasks.

4. CONCLUSION

In this paper, we proposed MSF-SER, a multi-granularity semantic fusion framework for speech emotion recognition. By introducing LES, GS, and ES, and integrating them through gated fusion and FM-MOE, our approach enhances acoustic modeling with enriched textual cues. Experiments on MSP-Podcast and IEMOCAP show consistent improvements over strong baselines, establishing MSF-SER as a top-performing system. In future work, we plan to extend this design to cross-lingual scenarios and explore its integration with additional modalities such as visual signals.

Acknowledgments

This work was supported in part by the Scientific Research Starting Foundation of Hangzhou Institute for Advanced Study (2024HIASC2001), in part by the National Natural Science Foundation of China (No. 62506091), in part by the Zhejiang Provincial Natural Science Foundation of China (No. LQN25F020001), and in part by the Key R&D Program of Zhejiang (2025C01104).

Use of Generative AI and AI-Assisted Tools. Language editing in throughout the manuscript was assisted by ChatGPT (OpenAI) to improve grammar and clarity; all scientific content was authored by the authors. During implementation, the authors used Cursor (an AI code assistant) for debugging support; no AI-generated code, figures, tables, or text were included in the manuscript. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the content.

Compliance with Ethical Standards

This study involved no human or animal subjects and did not require ethics approval.

References

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

[2] Mohammed Abdelwahab and Carlos Busso, "Study of dense network approaches for speech emotion recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5084–5088.

[3] Jiaqi Zhao, Fei Wang, Kun Li, et al., "Temporal-frequency state space duality: An efficient paradigm for speech emotion recognition," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[4] Aneesha Sampath, James Tavernor, and Emily Mower Provost, "Efficient finetuning for dimensional speech emotion recognition in the age of transformers," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[5] Hang Li, Wenbiao Ding, Zhongqin Wu, et al., "Learning fine-grained cross modality excitement for speech emotion recognition," *arXiv preprint arXiv:2010.12733*, 2020.

[6] Ankita Pasad, Chung-Ming Chien, Shane Settle, et al., "What do self-supervised speech models know about words?," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 372–391, 2024.

[7] Philip Lieberman, "Some acoustic correlates of word stress in american english," *The Journal of the Acoustical Society of America*, vol. 32, no. 4, pp. 451–454, 1960.

[8] Maureen de Seyssel, Antony D'Avirro, Adina Williams, et al., "Emphassess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models," *arXiv preprint arXiv:2312.14069*, 2023.

[9] Jingcheng Hu, Houyi Li, Yinmin Zhang, et al., "Multi-matrix factorization attention," *arXiv preprint arXiv:2412.19255*, 2024.

[10] Yinhan Liu, Myle Ott, Naman Goyal, et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[11] Alec Radford, Jong Wook Kim, Tao Xu, et al., "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[12] Ding Ding, Zeqian Ju, Yichong Leng, et al., "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[13] Ethan Perez, Florian Strub, Harm De Vries, et al., "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[14] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, et al., "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[15] Lucas Goncalves, Donita Robinson, Elizabeth Richerson, et al., "Bridging emotions across languages: Low rank adaptation for multilingual speech emotion recognition," in *Proc. Interspeech 2024*, 2024, pp. 4688–4692.

[16] Georgios Chochlakis, Turab Iqbal, Woo Hyun Kang, et al., "Modality-agnostic multimodal emotion recognition using a contrastive masked autoencoder," in *Proc. Interspeech 2025*, 2025, pp. 3005–3009.

[17] Wen Wu, Chao Zhang, and Philip C Woodland, "Estimating the uncertainty in emotion attributes using deep evidential regression," *arXiv preprint arXiv:2306.06760*, 2023.

[18] Minji Ryu, Ji-Hyeon Hur, Sung Heuk Kim, et al., "Pitch contour model (pcm) with transformer cross-attention for speech emotion recognition," in *Proc. Interspeech 2025*, 2025, pp. 4353–4357.

[19] Thanathai Lertpetchpun, Tiantian Feng, Dani Byrd, et al., "Developing a high-performance framework for speech emotion recognition in naturalistic conditions challenge for emotional attribute prediction," *arXiv preprint arXiv:2506.10930*, 2025.

[20] A Reddy Naini, Lucas Goncalves, Ali N Salman, et al., "The interspeech 2025 challenge on speech emotion recognition in naturalistic conditions," Interspeech, 2025.

[21] Pravin Mote, Berrak Sisman, and Carlos Busso, "Unsupervised domain adaptation for speech emotion recognition using k-nearest neighbors voice conversion," in *Proceedings of INTERSPEECH*, 2024.

[22] Abinay Reddy Naini, Mary A Kohler, Elizabeth Richerson, et al., "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12031–12035.