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

EmoJudge: LLM Based Post-Hoc Refinement for Multimodal Speech Emotion Recognition

Prabhav Singh[1], *Jesús Villalba*[1,2]

[1]Center for Language & Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]HLT Center of Excellence, Johns Hopkins University, Baltimore, MD, USA
{psingh54, jvillal7}@jhu.edu

Abstract

In SER, a significant challenge lies in building systems that can accurately interpret emotions in naturalistic conditions. To address this, we present EMOJUDGE, our submission to the SER in Naturalistic Conditions Challenge. For the categorical SER task, we propose a novel LLM-refined multimodal approach, while for the dimensional SER task, we propose a robust multimodal architecture. In both submissions, WavLM-Large is combined with attentive pooling aided by residual networks to extract acoustic features. For text, RoBERTa-Large captures linguistic nuances. Experimentation identifies late fusion with logistic regression as the optimal method for integrating modalities. For the categorical challenge, our novel contribution includes using transcripts, speaker indicators, and audio descriptions as input to an LLM for post-hoc correction of conflicting predictions. Results demonstrate improvements over the baseline in both tasks, highlighting the effectiveness of our proposed approach.

Index Terms: speech emotion recognition, large language models, multimodality, post-hoc refinement

1. Introduction

Speech Emotion Recognition (SER) aims to identify human emotions from speech signals, playing a crucial role in human-computer interaction. A significant challenge in SER is developing systems that accurately interpret emotions in naturalistic and spontaneous conditions. To address this, the Speech Emotion Recognition in Naturalistic Conditions Challenge[1] [1] at Interspeech 2025 provides a platform for researchers to develop and benchmark SER technologies using the MSP-Podcast corpus, which contains over 324 hours of naturalistic conversational speech [2].

A major challenge in identifying emotion in natural speech lies in the inherently noisy and context-dependent nature of vocal cues, the wide variability across speakers, and the scarcity of sufficiently large labeled datasets for training robust models [3, 4, 5]. These obstacles often lead to poor generalization, especially in real-world conditions where environmental factors and individual speaking styles introduce additional variability. Furthermore, the dataset in question, MSP-Podcast [6], contains short utterances with a low average word count. This further reduces the quality of uni-modal methods like textual embeddings.

We address these issues by proposing EMOJUDGE, a novel approach that uniquely combines foundational models with LLMs. Unlike existing LLM-based approaches that focus on evaluating the use of LLMs in refining the input features to an architecture [7], EMOJUDGE employs a *post-hoc refinement mechanism* that preserves the integrity of the primary recognition system while enhancing its performance. The framework first leverages established acoustic and textual models for initial predictions, then employs an LLM to refine these outputs by incorporating speaker information, prosodic features, and audio descriptions. The key contributions of our work for both tasks are:

1. *Categorical Emotion Recognition:* We employ late fusion of WavLM-Large [8] and RoBERTa-Large [9] predictions across eight emotion labels. We use a *weighted focal loss* to address the class imbalance challenge. An LLM is then fed with multimodal context and is used to refine the predictions.
2. *Emotional Attribute Prediction:* We employ the same architectural backbone as above for Arousal, Valence, and Dominance prediction. We utilize Concordance Correlation Loss [10] for optimal regression performance. LLM prompting is not used for post-hoc correction for this task. However, we observe improvements over the baseline on adding residual connections over the fully-connected layers and regularizing the attention mechanism.

Our approach marks a significant advancement over recent LLM applications in SER, which have primarily focused on emotional prompting for benchmarking [11] or transcript error reduction [7]. By introducing LLMs as a refinement mechanism rather than a feature modifier, we establish a novel approach for emotion recognition systems. The remainder of this paper is organized as follows: Section 2 reviews related work in SER, with emphasis on multimodal and LLM-based approaches. Section 3 describes the dataset and metrics. Section 4 details our proposed methodology, including model architecture and fusion strategies. Section 5 presents our experimental setup and training strategy. Section 6 presents our results for the challenge. We conclude with future research directions in Section 7.

2. Background and Related Work

Recent advancements in SER have increasingly leveraged multimodal methods, which integrate various data sources to enhance emotion classification accuracy. These methods typically combine audio signals with textual information, allowing for a more comprehensive understanding of emotional expressions. For instance, studies such as those by Hu et al. [12] propose multimodal multi-task learning frameworks that utilize dynamic fusion techniques to capture the nuances of emotional cues from both speech and text modalities, achieving state-of-the-art performance on benchmark datasets like IEMOCAP. The integration of different modalities not only improves recognition rates but also addresses the limitations inherent in single-modality

[1]https://lab-msp.com/MSP-Podcast_Competition/IS2025/

10.21437/Interspeech.2025-1141

approaches, such as noise and redundancy in feature extraction [13, 14].

Extensive research and experimentation have also been conducted to determine the best method to combine the modalities. While early fusion of embeddings was a trend that delivered decent results [15, 16], recent works have shown late fusion of modalities to be particularly effective [17, 18]. In parallel, applying Large Language Models (LLMs) in SER has emerged as a promising avenue. The recent research by Li et al. [11] explores different prompting mechanisms for speech emotion recognition. They report promising results on combining multiple dimensions of information into the prompt. Furthermore, approaches that translate speech characteristics into natural language descriptions have enabled LLMs to perform multimodal emotion analysis without architectural modifications [19].

3. Dataset and Metrics

This section introduces the dataset and defines the splits used for training and validation. We also discuss the metrics used to evaluate the experiments.

3.1. The MSP-Podcast Corpus

The MSP-Podcast corpus consists of spontaneous audio recordings from podcast segments. Each segment is annotated with categorical emotion labels: Angry, Sad, Happy, Surprise, Fear, Disgust, Contempt, and Neutral, and dimensional presence: Valence, Dominance, and Arousal. This study uses the default provided splits: *Train* and *Development*. The training set was used to build our models, while the development set served as an evaluation subset for hyperparameter tuning and validation. Table 1 shows the distribution of emotion categories in each split, along with the total number of files. We also evaluated our model on a held-out set taken from the development dataset (which we refer to as Dev-2). This split is kept unseen until the final evaluation.

Table 1: *Emotion distribution in the MSP-Podcast dataset across Train and Development splits.*

Emotion	Development (%)	Train (%)	Dev-2 (%)
Angry	23.11%	10.05%	12.50%
Sad	9.27%	9.41%	12.50%
Happy	25.12%	24.95%	12.50%
Surprise	3.91%	4.40%	12.50%
Fear	1.29%	1.67%	12.50%
Disgust	2.15%	2.14%	12.50%
Contempt	5.78%	3.72%	12.50%
Neutral	29.39%	43.65%	12.50%
Total Files	22,898	66,992	2,360

3.2. Metrics

The evaluation for Task 1 was conducted using the F1-Macro score, a metric that balances precision and recall across all classes, ensuring equal weight for each class regardless of its frequency. For Task 2, the Concordance Correlation Coefficient (CCC) was employed to evaluate the agreement between the predicted (y) and ground truth ($\hat{y}$) values. The formula for CCC is defined as

$$\text{CCC} = \frac{2 \cdot \rho \cdot \sigma_y \cdot \sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2},$$

where ρ is the Pearson correlation coefficient, σ_y and $\sigma_{\hat{y}}$ are the standard deviations, and μ_y and $\mu_{\hat{y}}$ are the means of y and $\hat{y}$, respectively.

4. Methodology

In this section, we present the proposed methodology. First, we discuss our choices for the foundational audio and text models. In the subsequent subsections, we discuss the preprocessing steps and architecture for both the audio and text modality, followed by the chosen fusion strategy. Finally, the LLM refinement is discussed with a focus on the used prompt engineering.

4.1. Foundational Model Selection

For the audio modality, WavLM-Large[2] was chosen as the foundational model for feature extraction due to its exceptional ability to capture long-range dependencies and robustness to noisy speech environments [20]. We also experimented with HuBERT [21], ECAPA-TDNN [22], and Wav2Vec2 [23]. For the text modality, we experimented with both RoBERTa [9] and DeBERTa [24]. We did not observe any observable difference in any particular architecture and hence settled on RoBERTa-Large[3] for further experimentation. To present an overview of the comparative performance of the foundational models, we report all scores in Table 2 for the validation set and also report test scores for submitted systems.

4.2. Normalization and Preprocessing

We process audio inputs at 16 kHz with a frame shift of 20 ms, aligning with WavLM's specifications. We kept the encoder's feature extractor frozen during training, allowing us to leverage the robust speech representations while fine-tuning the task-specific layers. We also normalized the audio samples using precomputed statistics (mean and standard deviation) on the train set. These statistics are then applied to validation and test files at inference time. We do not implement any specific preprocessing for the transcripts.

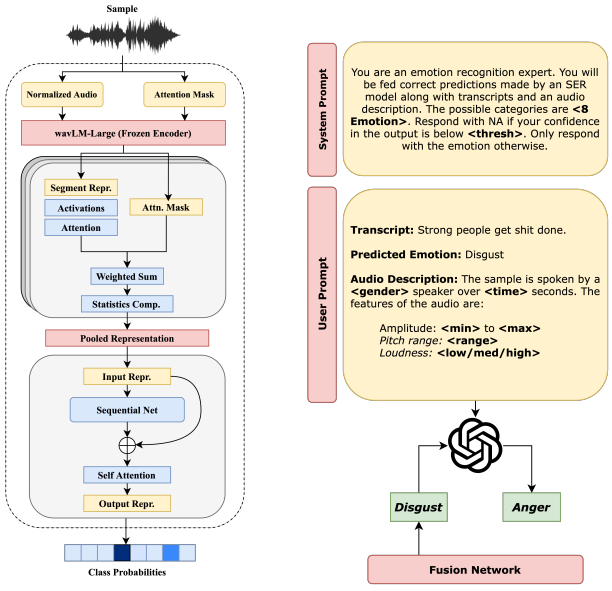
4.3. Audio Modality

Our audio processing framework focused on an architecture that addresses three key challenges: temporal dependency modeling, class imbalance, and effective feature aggregation from high-dimensional speech representations. A high-level overview of the architecture is shown in Figure 1a. The base architecture is taken from the baseline: Attentive Statistics Pooling [2], which addresses the variable-length nature of speech signals while preserving emotionally salient information. However, we supplement and build on this architecture by adding ℓ_2 regularization to the parameters responsible for computing the attention weights. Specifically, we regularize the attention parameter matrix $\mathbf{A}$ used in the attention computation to encourage stable and generalizable attention patterns. This attention formulation allows the model to dynamically weight different temporal regions based on their emotional content.

[2] https://huggingface.co/microsoft/wavlm-large

[3] https://huggingface.co/FacebookAI/roberta-large

The emotion classification/regression network consists of fully connected layers, each followed by layer normalization, a ReLU activation, and dropout. Residual connections are applied between the hidden layers to promote stable gradient flow and faster convergence.



(a) *Audio SER Framework* (b) *Prompting Framework*

Figure 1: *System Architectures*

4.4. Textual Modality

The textual analysis component leverages RoBERTa-Large [9], which consists of 24 transformer layers with 16 attention heads each. The model processes input text using RoBERTa's byte-level BPE tokenizer with a maximum sequence length of 128 tokens. For classification, we extract the [CLS] token representation (1024-dimensional embedding) from the final transformer layer, which serves as a comprehensive sentence embedding. This embedding is then passed through RoBERTa's classification head consisting of a dropout layer ($p = 0.3$) followed by linear layers that output emotion probabilities. No additional layers or pooling strategies were employed to maintain the model's pretrained characteristics while adapting to emotion recognition.

4.5. Fusion Framework

Our framework employed an L2-regularized logistic regression model to perform late fusion of the logits produced by audio and text modalities. This fusion strategy was made adaptable to both categorical emotion labels (Task 1) and dimensional emotion attributes (Task 2). The model's regularization strength, where $C = \frac{1}{\lambda}$ (with λ being the ℓ_2 regularization coefficient), was optimized through extensive grid search over $C \in [10^{-3}, 10^{3}]$, maximizing generalization performance.

4.6. LLM Refinement

For Task 1, the predicted emotion from the fusion network underwent a refinement step using OpenAI's GPT-4 model[4] [25]. As shown in Figure 1b, the LLM was prompted with three key components: the transcript, an audio description containing prosodic features (amplitude, pitch range, and loudness), and the predicted emotion. The audio description was constructed in text format by summarizing extracted features and including additional metadata such as the duration of the audio segment and speaking rate[5]. The LLM acted as an expert evaluator, assessing whether the predicted emotion aligned with both the semantic content and the described acoustic characteristics. To ensure computational efficiency, predictions were processed in batches using OpenAI's Batch API with a low-temperature setting of 0.3, maintaining response consistency.

A critical challenge observed during initial experiments was the LLM's tendency to misclassify samples when textual cues contradicted audio descriptors. To address this issue, we introduced a confidence threshold in the prompt design, instructing the LLM to modify a predicted label only if its confidence exceeded a predefined threshold. This step mitigates hallucination and ensures the robustness of the refinement process. Empirical evaluations indicated that setting the confidence threshold to 80% yielded optimal results. Table 2 demonstrates the efficacy of this approach, reporting scores with and without LLM refinement.

5. Training & Experimental Setup

In this section, we describe our approach to handling class imbalance and detail the training methodology and hyperparameter selection.

5.1. Loss Functions

For Task 1, we employed a weighted Focal Loss [26] to address class imbalance in both audio and text modalities. Focal Loss modifies the standard Cross-Entropy loss by down-weighting well-classified examples, enabling the model to focus on challenging samples. For a sample with true label y and predicted posterior probability p_y, the loss is defined as:

$$\mathcal{L}_{\text{focal}} = -\alpha(1 - p_y)^{\gamma} \log(p_y)$$

where α is the class-balancing factor and γ is the focusing parameter that determines the emphasis on misclassified examples. For Task 2, we optimize directly for the evaluation metric using CCC loss, computed as $\sum_{dim}(1 - CCC_{dim})$ across the three dimensions.

5.2. Training Protocol

We trained our audio framework using the AdamW optimizer [27] with a learning rate (LR) of 1e-5, implementing separate optimizers for the pooling layer and emotion regression network. To handle memory constraints while maintaining effective training, we used gradient accumulation with a batch size of 32 and accumulation steps of 4. Models were trained for 30 epochs, with checkpointing based on the F1-Macro score for Task 1 and CCC for Task 2. For the text modality, we trained using a mixed-precision strategy with gradient accumulation of 2 steps and a warmup period (10% of training steps). We again employed the AdamW optimizer with an LR of 1e-5 and weight decay of 0.01, training for 15 epochs with evaluation every 2000 steps based on the macro F1-score or CCC.[6]

[4] https://platform.openai.com/docs/models

[5] We use Librosa for computing all audio features.

[6] For all hyperparameters, we use the previously defined Dev-2 validation set.

Table 2: *Results for Task 1 and Task 2.* ^{L} *or* ^{S} *represent Large or Small version of the models. The final submitted system is marked with a* ✓. ***Test scores are only provided for systems submitted to the leaderboard.***

Task 1 (Categorical SER)					
Framework	**Fusion / Refinement**	**Validation Set**		**Test Set**	
		F1 Macro ↑	**Accuracy ↑**	**F1 Macro ↑**	**Accuracy ↑**
WavLM^{L} + RoBERTa^{L} ✓	Yes/Yes	**0.4292**	**0.4666**	**0.3388**	**0.3631**
WavLM^{L} + DeBERTa^{L}	Yes/Yes	0.3822	0.4353	0.3345	0.3603
WavLM^{L} + RoBERTa^{L}	Yes/No	0.3695	0.427	-	-
WavLM^{S} + RoBERTa^{S}	Yes/No	0.3224	0.415	0.2631	0.2984
Whisper^{S} + RoBERTa^{S}	Yes/No	0.3144	0.3939	-	-
ECAPP-TDNN + RoBERTa^{L}	Yes/No	0.2138	0.2968	0.1083	0.1325
Baseline		0.307	0.409	0.3293	0.3556
Task 2 (Dimensional Prediction)					
Framework	**Fusion**	**Validation Set**		**Test Set**	
		CCC (Val/Aro/Dom)	**Average CCC ↑**	**CCC (Val/Aro/Dom)**	**Average CCC ↑**
WavLM^{L} + RoBERTa^{L} ✓	Yes	**0.6891** / **0.6109** / 0.6679	**0.6560**	**0.6441** / **0.6229** / 0.4769	**0.5813**
WavLM^{L} + DeBERTa^{L}	Yes	0.6611 / 0.6102 / 0.6566	0.6426	-	-
Whisper^{S} + RoBERTa^{L}	Yes	0.6597 / 0.6032 / 0.6093	0.6241	0.5435 / 0.5693 / 0.4153	0.5094
WavLM^{L}	Audio Only	0.6641 / 0.595 / 0.6101	0.6231	0.6441 / 0.615 / 0.4527	0.5706
ECAPP-TDNN + RoBERTa^{L}	Yes	0.598 / 0.550 / 0.5970	0.5817	-	-
Baseline		0.652 / 0.579 / **0.688**	0.6396	0.6385 / 0.6232 / **0.4775**	0.5797

6. Results

This section presents a summary of results obtained from each system. All reported results are measured on the defined Development Set in Table 1.

6.1. Task 1: Categorical Emotion Recognition

Our experiments demonstrate the effectiveness of LLM-based refinement in categorical emotion recognition. The best-performing system (WavLM^{L} + RoBERTa^{L}) achieves an F1-Macro score of 0.4292 on the validation set, significantly outperforming the baseline (0.307). The impact of LLM refinement is particularly evident when comparing identical architectures with and without refinement (Row 1 vs Row 3). For instance, WavLM^{L} + RoBERTa^{L} shows a 6% improvement in F1-Macro (0.4292 vs 0.3695) when LLM refinement is applied. This improvement suggests that LLM refinement provides consistent benefits independent of the underlying architecture.

While more complex ensemble approaches might yield marginally better results, we focused on demonstrating the benefits of LLM refinement within a consistent architectural framework. The test set results (F1-Macro: 0.3388) achieve a performance advantage over the baseline (0.3293), while not adding the computational cost associated with larger and a higher number of models. We note that this improvement does not match the validation results due to the balanced test set distribution, which presents greater classification challenges compared to the more class-imbalanced validation set.

6.2. Task 2: Dimensional Prediction

For Task 2, which requires the prediction of dimensional attributes, the best results are again obtained with the multimodal fusion of WavLM^{L} and RoBERTa^{L} systems. On the validation set, this approach achieves an average CCC score of 0.6560 in comparison to a baseline of 0.6396. We specifically observe improvements in the valence dimension, suggesting that infusing textual cues might have a direct impact on modeling this dimension. We also note that the performance of an audio-only system (WavLM^{L}) remains competitive, especially for arousal, implying that lexical cues may play a less critical role for some emotional dimensions.

We also observe a pattern of test dominance scores being much lower than validation dominance scores. This is seen throughout rows 1-5 and also the baseline. We posit that this results from an inherent bias in the value distribution of dominance samples in the two sets. We also assume that this could be a result of confusion between arousal and dominance.

In summary, our results indicate that (i) LLM-based refinement consistently improves categorical emotion recognition, (ii) multimodal fusion provides benefits for dimensional emotion prediction—especially for valence and arousal—and (iii) further research is needed to address the inherent challenges in modeling the dominance dimension.

7. Conclusion & Limitations

In this work, we present EMOJUDGE as a submission to the SER in Naturalistic Conditions Challenge at Interspeech 2025. For Task 1, we present a novel LLM-based post-hoc refinement technique, enhancing prediction accuracy by integrating multimodal cues into LLM prompts. For Task 2, we utilized the same architecture and achieved competitive performance, especially for valence and arousal dimensions. Our system demonstrates robust performance on both tasks, outperforming baselines and aligning with the challenge's objectives to advance SER in naturalistic conditions. However, two limitations warrant discussion. First, the framework's performance on the dominance dimension lags behind other emotional attributes, suggesting that current architectures may not fully capture this aspect of emotional expression. Second, we note that we did not explore the possibility of using LLM Refinement for Task 2 and hope that future works explore the possibility of binning continuous values to allow the use of LLMs in the pipeline.

8. References

[1] A. Reddy Naini, L. Goncalves, A. N. Salman, P. Mote, I. R. Ülgen, T. Thebaud, L. Velazquez, L. P. Garcia, N. Dehak, B. Sisman, and C. Busso, "The interspeech 2025 challenge on speech emotion recognition in naturalistic conditions," in *Interspeech 2025*, vol. Under submission, Rotterdam, The Netherlands, August 2025.

[2] L. Goncalves, A. N. Salman, A. R. Naini, L. Moro-Velázquez, T. Thebaud, P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 247–254.

[3] C. Busso, M. Bulut, C. M. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[4] S. Poria, E. Cambria, A. Gelbukh, and F. Bisio, "Multi-modal sentiment analysis," *Cognitive Computation*, vol. 10, no. 3, pp. 487–499, 2018.

[5] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9–10, pp. 1062–1087, 2013.

[6] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:973607

[7] E. Zhang and C. Poellabauer, "Improving speech-based emotion recognition with contextual utterance analysis and llms," 2024. [Online]. Available: https://arxiv.org/abs/2410.20334

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "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. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2022.3188113

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[10] B. T. Atmaja and M. Akagi, "Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition," *Journal of Physics: Conference Series*, vol. 1896, no. 1, p. 012004, apr 2021. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1896/1/012004

[11] Y. Li, Y. Gong, C.-H. H. Yang, P. Bell, and C. Lai, "Revise, reason, and recognize: Llm-based emotion recognition via emotion-specific prompts and asr error correction," 2024. [Online]. Available: https://arxiv.org/abs/2409.15551

[12] H. Hu, J. Wei, H. Sun, C. Wang, and S. Tao, "Speech emotion recognition based on multimodal and multiscale feature fusion," *Signal, Image and Video Processing*, vol. 19, no. 2, Dec 2024.

[13] Z. Kang, J. Peng, J. Wang, and J. Xiao, "Speecheq: Speech emotion recognition based on multi-scale unified datasets and multitask learning," in *Interspeech*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250073154

[14] P. Singh, R. Srivastava, K. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Know.-Based Syst.*, vol. 229, no. C, Oct. 2021. [Online]. Available: https://doi.org/10.1016/j.knosys.2021.107316

[15] D. Prisayad, T. Fernando, S. Sridharan, S. Denman, and C. Fookes, "Dual memory fusion for multimodal speech emotion recognition," in *Interspeech*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260908088

[16] B. Bucur, I. Şomfelean, A. Ghiuruţan, C. Lemnaru, and M. Dînşoreanu, "An early fusion approach for multimodal emotion recognition using deep recurrent networks," in *2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2018, pp. 71–78.

[17] T. Thebaud, A. Favaro, Y. Guan, Y. Yang, P. Singh, J. Villalba, L. Mono-Velazquez, and N. Dehak, "Multimodal emotion recognition harnessing the complementarity of speech, language, and vision," in *Proceedings of the 26th International Conference on Multimodal Interaction*, ser. ICMI '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 684–689. [Online]. Available: https://doi.org/10.1145/3678957.3689332

[18] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition," *ArXiv*, vol. abs/2207.04697, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250426500

[19] Z. Wu, Z. Gong, L. Ai, P. Shi, K. Donbekci, and J. Hirschberg, "Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances," 2024. [Online]. Available: https://arxiv.org/abs/2407.21315

[20] D. Diatlova, A. Udalov, V. Shutov, and E. Spirin, "Adapting wavlm for speech emotion recognition," *ArXiv*, vol. abs/2405.04485, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269614106

[21] 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," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, Oct. 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.

[23] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[24] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *ArXiv*, vol. abs/2006.03654, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219531210

[25] O. J. Achiam and S. A. et al., "Gpt-4 technical report," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257532815

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53592270