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

ABHINAYA - A System for Speech Emotion Recognition In Naturalistic Conditions Challenge

Soumya Dutta[1], *Smruthi Balaji*[2], *Varada R*[1], *Viveka Salinamakki*[1], *Sriram Ganapathy*[1]

[1]Learning and Extraction of Acoustic Patterns (LEAP) Lab, Electrical Engineering, Indian Institute of Science, Bangalore, India
[2]Shiv Nadar University, Chennai, India
sriramg@iisc.ac.in

Abstract

Speech emotion recognition (SER) in naturalistic settings remains a challenge due to the intrinsic variability, diverse recording conditions, and class imbalance. As participants in the Interspeech Naturalistic SER Challenge which focused on these complexities, we present "Abhinaya", a system integrating speech-based, text-based, and speech-text models. Our approach fine-tunes self-supervised and speech large language models (SLLM) for speech representations, leverages large language models (LLM) for textual context, and employs speech-text modeling with an SLLM to capture nuanced emotional cues. To combat class imbalance, we apply tailored loss functions and generate categorical decisions through majority voting. Despite one model not being fully trained, the Abhinaya system ranked 4th among 166 submissions. Upon completion of training, it achieved state-of-the-art performance among published results, demonstrating the effectiveness of our approach for SER in real-world conditions.[1]

Index Terms: Speech Emotion Recognition (SER), Large Language Models (LLMs), Class Imbalance, Loss Functions.

1. Introduction

Speech carries significant para-linguistic information like cues about speaker, language, ambient conditions, emotions, and physiological/psychological health states of the speaker. The efforts to extract paralinguistic attributes has greatly benefited from evaluation campaigns and global benchmarking, much like other fields of machine learning. For example, speaker and language recognition evaluation challenges held over two decades [1, 2], speaker diarization challenges [3, 4], COVID-19 diagnostic challenges [5, 6], and speech recognition challenges [7, 8] have propelled the development of these technologies by fostering community participation.

On the task of speech emotion recognition, the conduct of challenges has been relatively sparse. One of the first such attempts for speech emotion recognition (SER), was held in 2009 [9], which has been extended to multimodal emotion recognition [10, 11]. Most widely used SER datasets are collected in controlled environments, featuring actors expressing scripted emotions [12, 13]. These datasets are typically small, and models trained on them struggle to generalize to real-world conditions. To address this limitation, the MSP-PODCAST corpus [14] was introduced as a large-scale, naturalistic dataset. The data formed the basis for the Odyssey 2024 emotion recognition challenge [15]. A new version of the MSP-PODCAST dataset is used for the Interspeech 2025 Emotion Recognition in Naturalistic Conditions Challenge [16].

The Naturalistic Conditions Challenge explored the following aspects: 1) Imbalanced training data with balanced test sets across different emotion classes, 2) High variability in emotional expression, featuring more than 2000 speakers, 3) Availability of text annotations for the training data, enabling the development of multi-modal (speech-text) systems [17].

This paper describes the details of "Abhinaya" (meaning - art of expression in *Sanskrit*) system, an SER model, which was developed as part of this challenge. The Abhinaya system directly addresses the challenge's key issues—class imbalance, multimodal modeling, and data variability—through an ensemble of speech-only, text-only, and speech-text joint models. Rather than building these components from scratch, we leverage SSL models (WavLM-Large [18]) and speech large language models (SLLMs) (SALMONN [19]) as speech feature encoders. Given the potential benefits of multimodal information, we transcribe speech using Whisper-large-v3 automatic speech recognition (ASR) system [20] and process the transcripts with LLMs. Further, we adopt an approach for the multimodal model, where speech and text are concatenated at the input and processed jointly using an SLLM. To mitigate class imbalance, we explore various loss functions for fine-tuning LLMs and SLLMs for SER. Overall, the Abhinaya system, based on an ensemble of five models, ranked 4th out of 166 entries on the categorical leaderboard[2]. Notably, one subsystem was not fully fine-tuned at the official submission deadline. After completion of this fine-tuning, Abhinaya achieved state-of-the-art (SoTA) performance in post-challenge evaluation.

The key contributions from this work are:

- Exploring Speech Large Language Models (SLLMs) as feature embedding models that can be fine-tuned for the task of emotion recognition from speech or speech-text jointly.
- Investigating specialized loss-functions for large model training in a class-imbalanced setting.
- Combining speech-only, text-only and speech-text models with zero-shot/fine-tuned settings using a majority voting rule, achieving SoTA results on the SER classification task in the naturalistic conditions challenge.

2. Background

Speech foundation models for SER: Speech foundation models trained in a self-supervised learning fashion such as wav2vec2.0 [21], HuBERT [22] or WavLM [18] yield rich speech representations for various tasks, including speech emotion recognition (SER). Both wav2vec2.0 and HuBERT embeddings have been used for SER [23, 24, 25]. The WavLM-large

[1]This work was partly funded by grants provided by British Telecom.

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

10.21437/Interspeech.2025-368

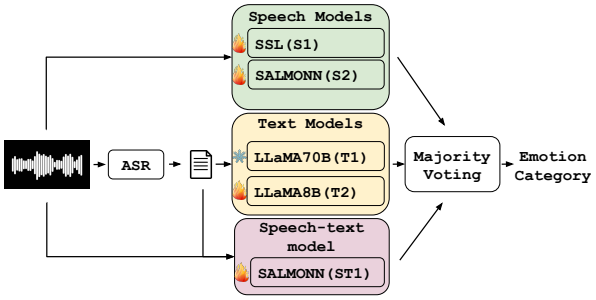

Figure 1: *Schematic of the different components of the Abhinaya SER system. We use three types of models - speech-only (**S1**, **S2**), text-only (**T1**, **T2**) and speech-text (**ST1**). Only **T1** is used in a zero-shot setting. The text used by the models are ASR transcripts generated by Whisper [20].*

model was fine-tuned on the MSP-PODCAST dataset [16], serving as the baseline system for this challenge.

Emotion recognition using LLMs: A comprehensive evaluation by Zhang et al. [26] highlights the effectiveness of LLMs in text sentiment classification. However, these models struggle with fine-grained emotion understanding, as shown recently by Dutta et al. [27]. Adapting text-based LLMs for speech tasks with speech-specific encoders has witnessed significant progress in recent years. Some examples of such speech large language models (SLLMs) are SALMONN [19], Qwen-Audio [28] and WavLLM [29]. The proposed Abhinaya system uses the SALMONN model for some of the sub-systems.

Loss Functions for imbalanced data: A common strategy to handle imbalanced datasets is using the weighted cross-entropy (WCE) loss, which utilizes sample-wise weights based on the class distribution. This was used in the challenge's baseline system [16]. Another approach, focal loss [30], dynamically re-weights samples based on prediction confidence. Both losses were explored in the Odyssey 2024 SER Challenge [15]. Additionally, we investigate vector scaling (VS) loss [31], which modifies the model's predictions using class-wise temperature scaling and bias correction.

3. Proposed SER System

The schematic of the Abhinaya system[3] is shown in Fig. 1.

3.1. Speech-only models

3.1.1. Wav-LM SSL (S1)

The speech foundation model used in this sub-system is the WavLM-large model [18]. The convolutional feature extractor is frozen, and transformer layers are fine-tuned. Since WavLM generates frame-level representations, we apply attentive pooling [32] to aggregate these features before emotion classification. The attentive pooling layer outputs a weighted mean of the features along with a weighted standard deviation, where the weights are learned with an attention mechanism.

3.1.2. SALMONN SLLM (S2)

The SALMONN speech large language model (SLLM) [19] consists of speech encoder - Whisper encoder [20] + BEATS encoder [33], with a text-based LLM (LLaMA [34]). The speech encoder and LLaMA weights remain frozen, while only the Q-former [35] and the low-rank adapters (LoRA) in the LLaMA module are trained for audio and speech tasks [19]. The SALMONN model is available in two sizes - 7B and 13B depending on the size of the back-end text LLM.

In our setting, we use all the layers of the SALMONN model except the final text-token layer. The representations from the final LLaMA transformer layer are passed through the attentive statistics pooling layer, and the model is fine-tuned for the emotion classification task with a softmax head. During fine-tuning, we update only the Q-Former and LoRA weights, along with the classification head.

3.2. Text-only models

In the Naturalistic Conditions Challenge, the reference text transcripts were provided only for the training and validation subsets, while the evaluation data lacked ground-truth transcripts. To ensure consistency, we used ASR transcripts generated by Whisper-Large v3 [20] for our model development during training and testing. We explore two choices for text-only models.

3.2.1. Zero-shot LLaMA (T1)

We use the `LLaMA-3.3-70B-Instruct` model [36] in zero shot setting for this task. The model was prompted to classify emotions into predefined categories.

3.2.2. Fine-tuned smaller text LLM (T2)

Prior studies [26, 27] indicate that LLMs often struggle with fine-grained emotion recognition in zero-shot scenario. To address this, we fine-tune the `LLaMA-3.1-8B` model [36] using LoRA for emotion recognition. Similar to the SALMONN fine-tuning setup, we extract last-layer representations from LLaMA, apply attentive statistics pooling, and feed the utterance-level embeddings to an emotion classification head.

3.3. Speech-text joint model (ST1)

Since speech large language models (SLLMs) primarily adapt LLMs for speech sequences, we explore their potential for joint speech-text modeling. For this purpose, we use the SALMONN-7B model. In addition to the speech sequence, we append the Whisper-generated text transcript before fine-tuning the SALMONN model using LoRA. As in speech/text only fine-tuning, we extract final-layer representations, apply attentive statistics pooling, and train the classification head.

3.4. Loss functions

The dataset is highly imbalanced with the frequency of the majority class ("neutral") almost 26 times that of the minority class ("fear") in the training data. To address this issue, we explore various loss functions:

3.4.1. Weighted cross entropy (WCE)

Denoting the total number of data points by N, the number of classes by C, the weight assigned for a particular class, $c \in [1, 2, \ldots, C]$, is given by $w^c = \frac{N}{N_c \times C}$, where there are N_c samples belonging to class c in the training data. The cross-entropy loss l_i for sample i is given by:

$$l_i = -\sum_{c=1}^{C} t_{ic} w^c \log(p_{ic}) \quad ; \quad \mathbb{E} = \frac{1}{N} \sum_{i=1}^{N} l_i \qquad (1)$$

where p_{ic} is the prediction probability of class c for sample i. t_{ic} is set to 1 for the true class c of sample i, else it is kept at 0. The loss $\mathbb{E}$ is defined as the weighted cross-entropy loss (WCE).

[3] Code:`https://github.com/iiscleap/ABHINAYA`

3.4.2. Weighted FoCal Loss (WFL)

To further emphasize hard-to-classify samples, we explore the focal loss [30], which adjusts the weight based on the model's confidence. The weighted focal loss (WFL) is defined as:

$$l_i = -\sum_{c=1}^{C} t_{ic} w^c (1 - p_{ic})^{\gamma} \log(p_{ic}), \quad (2)$$

and the total loss $\mathbb{E}$ is the same as in Eq. 1. Here, $\gamma \geq 0$ is a hyper-parameter.

3.4.3. Vector scaling (VS) loss

The vector scaling (VS) loss adjusts predictions using class-dependent temperature scaling and bias correction. It modifies the pre-softmax model prediction, z_{ic}, as:

$$\hat{z}_{ic} = (\frac{N_c}{N_{max}})^{\gamma} z_{ic} + \tau \log(\frac{N_c}{N}) \quad (3)$$

where $\gamma \geq 0$ and $\tau \geq 0$ are hyper-parameters, while $N_{max} = \max_{c \in [1,2,\ldots,C]} N_c$. The softmax $\hat{p}_{ic} = softmax(\hat{z}_{ic})$ is used in the weighted loss:

$$l_i = -\sum_{c=1}^{C} t_{ic} w^c \log(\hat{p}_{ic}) \quad (4)$$

where w^c and t_{ic} are defined as before. The total loss ($\mathbb{E}$) is the same as in Eq. 1.

4. Experiments

4.1. Dataset

The MSP-PODCAST [14] dataset serves as the basis for this challenge. The training, validation and the test splits contain 84260, 31961 and 3200 speech files respectively. The dataset includes annotations for 8 primary emotion categories: "happy", "angry", "sad", "neutral", "surprise", "fear", "contempt" and "disgust". Additionally, two other labels—"other" (denoting emotions outside the predefined categories) and "X" (cases with no annotator consensus)—are present but excluded from emotion modeling. The training and the validation distribution for the 8 categories are reported in Naini et al. [16]. For the model selection, we construct a balanced validation set, using the same number of utterances (326) for each of the 8 emotion classes. The test data labels in the challenge were not provided and the performance of the models was only available through the leaderboard.

4.2. Implementation Details

Except for T1, which operates in a zero-shot manner, all models are fine-tuned using the AdamW optimizer [37] with a learning rate of $1e$-5. For models utilizing LoRA (S2, T2, and ST1), the LoRA parameters [38] are set as $r = 8$ and $\alpha = 32$, with a dropout rate of 0.1. Training configurations vary across models due to the diverse memory requirements:

- For S1 and T2, the batch size is set to 8 with gradient accumulation over 4 steps.
- For SALMONN-based models (S2 and ST1), the batch size is set to 4 with gradient accumulation over 8 steps.

In speech-only models (S1 and S2) and the speech-text model (ST1), audio samples are processed with a maximum duration of 10 seconds. All four models are trained for 20 epochs, with the checkpoint yielding the highest validation macro-F1 score selected for evaluation. Regarding loss functions, models trained with WFL (Eq. 2) use $\gamma = 2$, while those trained with VS loss (Eq. 3) have hyper-parameters $\gamma = 0.3$ and $\tau = 1$. Once the outputs from the five models are available, majority voting is applied across the five classifiers, with S2 serving as the tiebreaker due to its superior validation performance.

Table 1: *The performance (in %) of the different models on the balanced validation and test data. The performance of certain models on the test data are not reported as they were not submitted to the leaderboard. $*$ indicates the test score evaluated by the organizers after the official challenge deadline.*

Models	Modality	Loss	Val. F1 (macro)	Test F1 (macro)
Baseline [16]	Speech	WCE	-	32.93
S1 (WavLM-large-317M)	Speech	WFL	34.43	33
S2 (SALMONN-13B)	Speech	WFL	37.68	35.34
T1 (LLaMA-3.3-70B)	Text	-	32.78	-
T2 (LLaMA-3.1-8B)	Text	VS	33.68	-
ST1 (SALMONN-7B)	Speech-text	VS	35.43	-
ABHINAYA (4th place)	Speech-text	-	**43.01**	41.81
ABHINAYA* (SoTA)	Speech-text	-	42.31	**44.02**

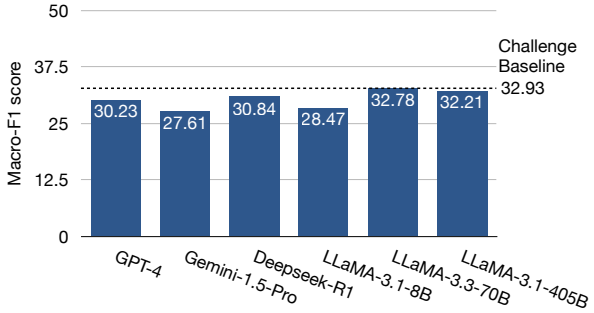

Figure 2: *Validation macro F1-score (in %) for different LLMs [36, 39, 40, 41] evaluated in zero-shot setting using the ASR transcripts. The LLaMA models considered are the Instruct versions. The baseline performance is also shown.*

4.3. Performance of Abhinaya

Table 1 presents the validation and test set results. The key findings are summarized below:

- **Fine-tuned SALMONN-13B (S2) achieves the best individual system performance**: Note that, its zero-shot validation performance is only 18.63%, hence fine-tuning significantly improves the performance.
- **Fine-tuning a small model versus zero-shot evaluation of a large model**: The fine-tuned LLaMA-3.1-8B (T2) model outperforms the zero-shot LLaMA-3.3-70B model (T1).
- **Ensembling improves performance**: A majority vote ensemble of the five models achieves a 24.56% relative gain over the best individual model (S2) on the test set.
- **Significant improvement over the challenge baseline**: Our system achieves a 33.68% relative improvement over the baseline, demonstrating its effectiveness for naturalistic SER.
- **Post-challenge evaluation confirms state-of-the-art performance**: At the challenge submission deadline, the ST1 fine-tuning had completed only 3 epochs. We continued this

Table 2: *The performance of the models with different loss functions. All results are macro F1-scores (%) on the balanced validation data.* ***Bold*** *indicates best loss function for each model.*

Models	WCE	WFL	VS
S1	33.07	**34.43**	32.12
S2	36.34	**37.68**	33.17
T2	29.79	30.12	**33.68**
ST1	33.92	34.73	**35.43**

model fine-tuning till completion (20 epochs) and the organizers performed a post-challenge evaluation of the Abhinaya system with this update. The final system performance of 44.02% surpasses the best published result in the challenge, thereby providing the SoTA results on this task.

4.4. Evaluation of different LLMs for text-only models

For the T1 zero-shot setting, we evaluate multiple large language models (LLMs) for the task of emotion recognition using a same prompt structure. The comparative performance on the validation set is presented in Fig.2. The key observations are:

- LLaMA-3.3-70B-Instruct achieves the best performance, outperforming all other LLMs considered here.
- While the zero-shot performance of LLaMA-3.1-8B-Instruct model is only 28.47%, the fine-tuning with VS loss improves the performance to 33.68% (Table 1).
- The other "larger" LLMs did not perform as well on the emotion recognition task with ASR transcripts.

4.5. Choice of loss functions

The impact of different loss functions is analyzed in Table 2. The key observations are:

- Speech-only models (S1, S2) benefit most from weighted focal loss, suggesting its effectiveness in handling class imbalance for SER. In contrast, VS loss is effective for the text-based (T2) and speech-text (ST1) models.
- The stronger zero-shot performance of T2 (28.47% versus 18.63% for S2) suggests better initial separation between emotion classes. We hypothesize that this initial separation aids VS loss, which adjusts logits class-wise, whereas speech models with less discriminative initial representations benefit more from sample-wise loss weighting.

4.6. Speech-text versus speech-only SLLMs

To assess the impact of joint speech-text fine-tuning, we fine-tune a SALMONN-7B model using speech inputs. This setup mirrors model S2 but replaces the SALMONN-13B model with the smaller 7B variant for a direct comparison with ST1 model, which incorporates both speech and text. The speech-only model achieves 33.87% validation macro-F1, while ST1 achieves 35.43%, demonstrating the benefits of multimodal fine-tuning. However, speech-text models incur higher computational costs due to the longer input sequences.

4.7. Class-wise performance analysis

Table 3 presents macro-F1 scores per class for Abhinaya and its components. The speech models (S1, S2) perform poorly on the least frequent classes [16]: "fear", "contempt", "disgust". In contrast, the fine-tuned text model (T2) and the speech-text model (ST1) show improved performance on these classes. Interestingly, despite its weaker overall performance (Table 1), the zero-shot text model (T1) outperforms the speech-based models on the three rare classes and even outperforms the ensemble for the least frequent class, "fear". The speech-text model (ST1) generally performs between the speech and text models across classes (except "neutral"), while the ensemble achieves the best scores in four out of eight emotion categories.

Table 3: *Class-wise macro F1-scores (%) for the balanced validation data for Abhinaya and the different model components.* ***Bold*** *indicates best model for each emotion category.*

Emotion Classes	S1	S2	T1	T2	ST1	ABHINAYA
Angry	49.74	50.60	32.48	37.13	50.73	**51.49**
Contempt	20.02	13.74	22.37	22.56	**25.62**	22.87
Disgust	18.91	31.53	33.33	**37.31**	32.08	36.13
Fear	12.22	16.79	**29.22**	27.74	29.09	26.41
Happy	55.93	**62.94**	40.00	43.33	50.88	62.83
Neutral	33.13	37.06	31.24	20.45	16.37	**40.29**
Sad	50.60	51.47	40.69	47.47	46.86	**55.62**
Surprise	34.94	37.31	32.93	33.48	31.83	**42.81**

Table 4: *Results (in %) with majority voting from different combinations of models from Abhinaya.*

Model ensemble	S1	S2	T1	T2	ST1	Val. F1 (macro)	Test F1 (macro)
Comb. I	✗	✓	✗	✓	✓	40.36	39.74
Comb. II	✓	✗	✗	✓	✓	39.07	–
Comb. III	✓	✓	✗	✓	✓	40.99	–
Comb. IV	✓	✓	✓	✓	✗	40.48	–
ABHINAYA	✓	✓	✓	✓	✓	42.31	44.02

4.8. Majority voting with other combinations

We evaluate model combinations using majority voting (Table 4). The best-performing trio—S2 (speech), T2 (text), and ST1 (speech-text)—achieves 40.36% validation macro-F1 and 39.74% on the test set (Comb. I). Adding S1 improves validation performance by 0.63% absolute (Comb. III), while ensembling all five models boosts it by nearly 1.5% absolute (ABHINAYA). Since T1, T2, and ST1 perform better on the minority classes, ensembling all models enhances performance on these underrepresented categories compared to ensembling only S1, S2, T2, and ST1. The impact of ST1 is evident from the nearly 2% absolute drop in Comb. IV when it is removed. We also evaluate majority voting with the three models having the fewest parameters (S1, T2, and ST1) (Comb. II), using ST1 for tie-breaking. The validation F1 score drops to 39.07%, highlighting the need for larger models for the SER task.

5. Conclusion

In this paper, we present Abhinaya, a system for speech emotion recognition (SER) developed as part of the Interspeech Naturalistic Speech Emotion 2025 challenge. Our approach combines five models leveraging large language models (LLMs) and speech large language models (SLLMs)—two speech-based, two text-based, and one multimodal model. To address class imbalance, we explore various loss functions, demonstrating their impact on model performance. Our model ensemble achieves the best published results on this task. Experimental analysis highlights the importance of the different components and loss functions suitable for fine-tuning each model.

6. References

[1] C. S. Greenberg *et al.*, "Two decades of speaker recognition evaluation at the national institute of standards and technology," *Computer Speech & Language*, vol. 60, p. 101032, 2020.

[2] Y. Lee *et al.*, "The 2022 NIST Language Recognition Evaluation," in *Interspeech 2023*, 2023, pp. 1928–1932.

[3] N. Ryant *et al.*, "The Third DIHARD Diarization Challenge," in *Interspeech 2021*, 2021, pp. 3570–3574.

[4] S. Baghel *et al.*, "Summary of the DISPLACE challenge 2023-DIarization of SPeaker and LAnguage in Conversational Environments," *Speech Communication*, vol. 161, p. 103080, 2024.

[5] A. Muguli *et al.*, "DiCOVA Challenge: Dataset, Task, and Baseline System for COVID-19 Diagnosis Using Acoustics," in *Interspeech 2021*, 2021, pp. 901–905.

[6] N. K. Sharma *et al.*, "The second DICOVA challenge: Dataset and performance analysis for diagnosis of COVID-19 using acoustics," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 556–560.

[7] M. K. Nandwana *et al.*, "The VOiCES from a Distance Challenge 2019," in *Interspeech 2019*, 2019, pp. 2438–2442.

[8] S. Watanabe *et al.*, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. CHiME 2020*, 2020, pp. 1–7.

[9] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Interspeech 2009*, 2009, pp. 312–315.

[10] B. Schuller *et al.*, "Avec 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Proceedings, Part II*. Springer, 2011, pp. 415–424.

[11] L. Stappen *et al.*, "MUSE 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, 2020, pp. 35–44.

[12] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *LREC*, vol. 42, no. 4, pp. 335–359, 2008.

[13] ——, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[14] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[15] L. Goncalves *et al.*, "Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 247–254.

[16] A. Reddy Naini *et al.*, "The Interspeech 2025 Challenge on Speech Emotion Recognition in Naturalistic Conditions," in *Interspeech 2025*, vol. Under submission, Rotterdam, The Netherlands, August 2025.

[17] S. Dutta and S. Ganapathy, "Leveraging content and acoustic representations for efficient speech emotion recognition," *arXiv preprint arXiv:2409.05566*, 2024.

[18] S. Chen *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.

[19] C. Tang *et al.*, "SALMONN: Towards Generic Hearing Abilities for Large Language Models," in *The Twelfth International Conference on Learning Representations*, 2024.

[20] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[21] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[22] W.-N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[23] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Interspeech*, pp. 3400–3404, 2021.

[24] S. Dutta and S. Ganapathy, "HCAM–Hierarchical Cross Attention Model for Multi-modal Emotion Recognition," *arXiv preprint arXiv:2304.06910*, 2023.

[25] M. Pastor *et al.*, "Cross-corpus speech emotion recognition with hubert self-supervised representation," in *INTERSPEECH 2022*, 2022, pp. 76–80.

[26] W. Zhang *et al.*, "Sentiment Analysis in the Era of Large Language Models: A Reality Check," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 3881–3906.

[27] S. Dutta and S. Ganapathy, "LLM supervised Pre-training for Multimodal Emotion Recognition in Conversations," *arXiv preprint arXiv:2501.11468*, 2025.

[28] Y. Chu *et al.*, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[29] S. Hu *et al.*, "WavLLM: Towards robust and adaptive speech large language model," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 4552–4572.

[30] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[31] G. R. Kini *et al.*, "Label-imbalanced and group-sensitive classification under overparameterization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 970–18 983, 2021.

[32] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech 2018*, 2018, pp. 2252–2256.

[33] S. Chen *et al.*, "BEATs: Audio Pre-Training with Acoustic Tokenizers," in *International Conference on Machine Learning*. PMLR, 2023, pp. 5178–5193.

[34] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[35] J. Li *et al.*, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 19 730–19 742.

[36] A. Dubey *et al.*, "The LLAMA 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR,*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[38] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[39] D. Guo *et al.*, "DeepSeek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[40] G. Team *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[41] J. Achiam *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.