The Interspeech 2025 Challenge on Speech Emotion Recognition in Naturalistic Conditions

Abinay Reddy Naini[1,2], *Lucas Goncalves*[2], *Ali N. Salman*[2], *Pravin Mote*[1,2], *Ismail R. Ülgen*[3], *Thomas Thebaud*[3], *Laureano Moro-Velazquez*[3], *Leibny Paola Garcia*[3], *Najim Dehak*[3], *Berrak Sisman*[3], *Carlos Busso*[1]

[1]Language Technologies Institute (LTI), Carnegie Mellon University, USA
[2]The University of Texas at Dallas, USA [3]CLSP, Johns Hopkins University, USA

Abstract

The Interspeech 2025 *speech emotion recognition in naturalistic conditions* challenge builds on previous efforts to advance *speech emotion recognition* (SER) in real-world scenarios. The focus is on recognizing emotions from spontaneous speech, moving beyond controlled datasets. It provides a framework for speaker-independent training, development, and evaluation, with annotations for both categorical and dimensional tasks. The challenge attracted 93 research teams, whose models significantly improved state-of-the-art results over competitive baselines. This paper summarizes the challenge, focusing on the key outcomes. We analyze top-performing methods, emerging trends, and innovative directions. We highlight the effectiveness of combining foundational models based on audio and text to achieve robust SER systems. The competition website, with leaderboards, baseline code, and instructions, is available at: `https://lab-msp.com/MSP-Podcast_Competition/IS2025/`.

Index Terms: Speech emotion recognition, human-computer interaction, speech challenge

1. Introduction

Speech emotion recognition (SER) is crucial to enabling technologies that can interpret and react to human emotions [1–4]. Increasing the robustness of SER systems can facilitate the deployment of real-world SER applications, from mental health monitoring to human-computer interaction [5, 6]. To drive innovation in this field, we organized the *speech emotion recognition in naturalistic conditions* challenge as part of Interspeech 2025, building on the success of the Odyssey 2024 SER Challenge [7]. Unlike traditional datasets derived from acted or semi-structured scenarios, which often suffer from exaggerated emotions and limited spontaneity [8,9], this challenge leverages the MSP-Podcast corpus [10]. This database offers a rich collection of naturalistic, diverse, and well-annotated speech samples, capturing the subtlety and variability of authentic human emotions [11]. By focusing on natural emotional recordings, we aim to address the limitations of conventional datasets and foster the development of SER systems that perform reliably in real-world conditions.

The Odyssey 2024 SER Challenge [7] successfully demonstrated the potential of leveraging naturalistic datasets such as the MSP-Podcast corpus [10] to advance SER research, with 31 teams contributing with innovative methodologies. However, several gaps remained, particularly in exploring the integration of *state-of-the-art* (SOTA) speech and text-based foundational models, which have recently shown groundbreaking performance in various domains. For the Interspeech 2025 challenge, 93 teams participated, reflecting the growing interest in SER. The success of this challenge also provided an opportunity to analyze a wider range of techniques, including the impact of recent advances in text-based foundational models and *large language models* (LLMs), which have revolutionized various domains in natural language processing and speech technology. The challenge established SOTA performance for classifying emotional categories (task 1; eight class problems for anger, sadness, happiness, fear, neutral, contempt, disgust, and surprise) and predicting emotional attributes (task 2; arousal, valence, and dominance). The diverse range of approaches explored in this challenge, supported by the availability of an expanded training set from the MSP-Podcast corpus, offers valuable guidance on the future directions of SER, illustrating the techniques that lead to consistent improvements.

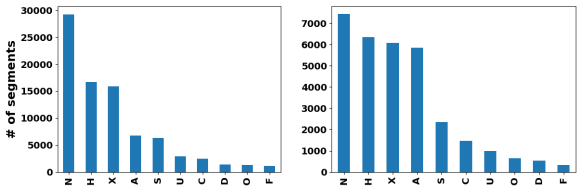
After summarizing the challenge's tasks, this paper discusses the lessons learned by exploring deeper insights from the methodologies driving SOTA performance. Our analysis is based on the information provided by the research teams on their systems that outperformed our competitive baselines, describing their approaches, training processes, and the advantages and limitations of their methods. Our analysis revealed a consistent trend: the complementary nature of speech and text-based foundational models. A multimodal approach that combined speech and text was pivotal in achieving the highest results. Other successful strategies include innovative fine-tuning and ensemble techniques, which led to significant performance improvements. These findings highlight the effectiveness of multimodal frameworks and pave the way for practical applications with reliable SER systems. This paper presents a comprehensive analysis of the challenge outcomes, identifying emerging trends and open questions that will drive continued progress in the field.

2. Description of the Challenge

2.1. The MSP-Podcast Corpus

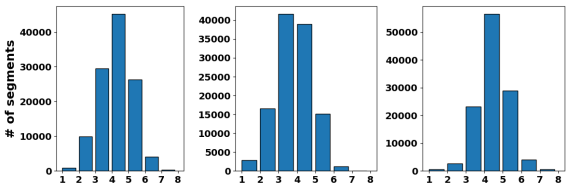
The challenge utilizes a subset of release 1.12 of the MSP-Podcast corpus [10]. This corpus is a rich collection of spontaneous emotional speech from diverse podcast recordings. Each speaking turn is annotated by at least five annotators for categorical emotions (anger, sadness, happiness, surprise, fear, disgust, contempt, neutral, and "other") and the emotional attributes of arousal (calm to active), valence (negative to positive), and dominance (weak to strong) [12]. The ground truth labels for categorical emotions are determined through a plurality vote, ensuring that the consensus emotion reflects the most consistent interpretation. The emotional attributes are rated on a seven-point Likert scale, with the consensus score being the average of the annotators' ratings.

The dataset also includes human-labeled transcripts and



(a) Training Set (b) Development Set

Figure 1: *Distribution of categorical emotions across training and development sets. A = Anger, C = Contempt, D = Disgust, F = Fear, H = Happiness, N = Neutral, O = Other, S = Sadness, U = Surprise, X = No agreement.*



(a) Arousal (b) Valence (c) Dominance

Figure 2: *Overall emotional attributes distribution for the training and the development set combined.*

force-aligned text-audio mappings to support multimodal approaches. The dataset is partitioned into speaker-independent training, development, and test sets to ensure fair evaluation. For this challenge, the training (2,114 speakers - 84,260 segments) and development (714 speakers, 31,961 segments) sets include samples in release 1.12 with speaker information. Figure 1 presents the distribution of categorical emotions in the training and development sets, highlighting the dataset's diversity. Figure 2 shows the distribution of these attributes across training and development sets, reflecting the natural variability in spontaneous speech. The test set corresponds to the *test 3* set in the corpus, which contains 3,200 segments (speaking turns) from 256 speakers. The key difference is that this set is perfectly balanced across primary categorical emotions (i.e., 400 sentences for each of the eight emotions, excluding "other (O)" and the segments with no agreement (X)). This structured partitioning ensures reliable evaluations, enabling models to be tested on both categorical and dimensional tasks for emotion recognition.

2.2. Challenge Tracks

This challenge has two core tracks: (1) categorical emotion recognition and (2) emotional attribute prediction.

Categorical emotion recognition: This task focuses on classifying speech samples into one of eight distinct emotional categories: anger, happiness, sadness, fear, surprise, contempt, disgust, and neutral. The performance in this track is evaluated using the Macro-F1 score, providing a robust measure of classification accuracy across all categories. As aforementioned, the evaluation test set is balanced across emotions.

Emotional attributes prediction: This task focuses on predicting a score for emotional attributes, including arousal, valence, and dominance. These emotional descriptors capture the subtle variations in emotional expression, enabling a more nuanced understanding of emotional states. Participants are tasked with predicting these continuous values, with performance evaluated using the *concordance correlation coefficient* (CCC) [13] to assess the alignment between predicted and ground truth values. We sort the results in the challenge by taking the average CCC for arousal, valence, and dominance.

We only released the speech files of the test set to maintain integrity in the challenge, withholding the emotional labels, transcriptions, and speaker information. The participants were asked to submit their predictions through a secure evaluation platform. The results are returned to the teams and automatically uploaded to the leaderboard. The teams were allowed to submit one prediction per week per task, except for the last week, where they were allowed two predictions per task.

3. Baseline

Our baseline model for both tasks follows a similar strategy to the one used in the Odyssey 2024 SER Challenge [7], but leverages an expanded training set from the MSP-Podcast corpus, which was provided for this challenge. The model consists of a *fine-tuning* module built upon the WavLM-large *self-supervised learning* model [14,15], followed by a prediction head comprising attentive statistics pooling [16] and *fully connected* layers. The pooling layer assigns adaptive weights to different frames, computing weighted means and standard deviations, which are fed into the FC layers for final predictions.

3.1. Implementation Details

The WavLM-large [17] model consists of 24 transformer layers and approximately 310M parameters. We fine-tune the model using pre-trained weights from the Hugging Face repository [18]. The model is trained for 30 epochs with a batch size of 32, using the Adam optimizer with a learning rate of 1e-5.

For emotion attribute prediction, separate models are trained for arousal, valence, and dominance, minimizing the loss function $\mathcal{L}_{CCC} = 1 - CCC$:

$$\mathcal{L}_{CCC} = 1 - \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

where μ_x and μ_y represent the means of the actual and predicted scores, respectively, σ_x and σ_y denote their standard deviations, and ρ is the Pearson's correlation coefficient.

For categorical emotion classification, the model is trained using the *cross-entropy* (CE) loss:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{M} w_c t_{o,c} \log(p_{o,c}) \quad (2)$$

where M represents the number of classes, c is the correct label for observation o, $p_{o,c}$ is the predicted probability that observation o belongs to class c, $t_{o,c}$ is 1 when observation o belongs to class c, and 0 otherwise, and w_c is the weight assigned

Table 1: *Baseline performance on categorical emotion recognition and emotional attributes predictions.*

Categorical Emotion Recognition

Model	F1-Macro	F1-Micro
Baseline	0.329	0.355

Emotional Attributes Prediction

Model	Arousal	Valence	Dominance	Average
Baseline	0.623	0.638	0.477	0.579

Table 2: *Summary of Top-Performing Teams' Methods and Techniques*

	Team Name	Speech Foundational Models	Other Foundational Models	Losses & Metrics	Class Imbalance	F1 Macro
Task 1	NTUA [19]	WavLM, Whisper, HuBERT	RoBERTa, ModernBERT	CrossEntropy, F1	Soft Labels	0.4316
	SAIL [20]	Whisper	RoBERTa	KLD Loss	Data Aug	0.4281
	ABHINAYA [21]	WavLM, SALMONN	LLaMA-3	Focal Loss	Weighted Loss	0.4181
	Voinosis [22]	WavLM, Whisper, HuBERT	BERT, T5	CrossEntropy	Class Weights	0.4101
	UNICAMP [23]	WavLM, Whisper, HuBERT	RoBERTa, DeBERTa	Weighted-CE, Rank	Batch Balancing	0.4094
	NU [24]	Whisper, HuBERT	RoBERTa, GPT-4	CrossEntropy	Weighted Loss	0.4033
	BSC-UPC [25]	Wav2Vec2, WavLM, Whisper	RoBERTa, DeBERTa	F1, Focal Loss	Weighted Loss	0.4006
	SRPOL [26]	WavLM, Whisper	RoBERTa	Weighted CE	Weighted Loss	0.3784
	Team Name	**Speech Foundational Models**	**Other Foundational Models**	**Losses & Metrics**	**Multi-Task**	**CCC Avg**
Task 2	SAIL [27]	WavLM, Whisper	RoBERTa	1-CCC	Yes	0.6076
	SRPOL [26]	WavLM, Whisper	RoBERTa	BSE, MSE Losses	Yes	0.6003
	SEU_AIPLab [28]	Wav2Vec2, Whisper, HuBERT	-	1-CCC	No	0.5955
	Voinosis [22]	WavLM, Whisper, HuBERT	BERT, T5	1-CCC	No	0.5928
	ABHINAYA [21]	WavLM, SALMONN-13B	-	1-CCC	No	0.5871

to class c, reflecting its inverse frequency. This weighting approach helps mitigate class imbalance by giving higher importance to underrepresented emotion classes.

This baseline serves as a benchmark for evaluating the effectiveness of different modeling approaches explored by participants in the challenge.

3.2. Baseline Results

Table 1 presents the baseline results for both tasks on the test set. For categorical emotion recognition across eight emotion classes, the model achieves an F1-micro score of 0.355 and an F1-macro score of 0.329. For emotional attribute prediction, the model obtains CCC scores of 0.623 for arousal, 0.638 for valence, and 0.477 for dominance. The average CCC across emotional attributes is 0.579. The lower performance on dominance prediction likely stems from its distribution (Fig. 2), where extreme dominance values are underrepresented, making it more challenging for the model to learn robust relevant patterns.

4. Observations and Findings

Ninety-three research teams participated in the challenge, submitting 166 entries to task 1 (categories) and 111 to task 2 (attributes). Twenty-eight teams submitted classification results for task 1 that were better than the baseline, whereas ten teams submitted predictions for task 2 that were better than the corresponding baseline. Table 2 lists the top-performing teams. This section analyzes and discusses the results and information provided by the teams that outperformed the baselines of the corresponding tasks and responded to our surveys (17 for task 1, 8 for task 2). Our analysis primarily focuses on the categorical task, given that similar architecture choices were used for the emotional attribute prediction task with slight variations in loss functions and training strategies.

Foundational models: Most top-performing teams employed multimodal frameworks, combining speech and text-based foundational models [15, 29, 30] (76%). For speech encoding, WavLM [17] (70%) and Whisper [31] (47%) were the most popular choices, while RoBERTa [32] (53%) and LLaMA (24%) were frequently used for text-based representations. In addition to these commonly used models, several teams explored advanced foundational models such as LLaMA-3 [33] (24%), GPT-4 [34] (18%), and ModernBERT [35] (5%), underscoring the growing trend of leveraging diverse multimodal architectures for robust emotion recognition. Other multimodal approaches demonstrated the growing use of cross-modal attention mechanisms to better integrate speech and text. The UNICAMP team extended conventional architectures by combining audio, text, and paralinguistic features through cross-attention blocks, enabling richer representations. ABHINAYA explored fine-tuning SALMONN [36] for SER tasks and combined it with speech-text fusion using LLaMA-3 [33] to boost accuracy.

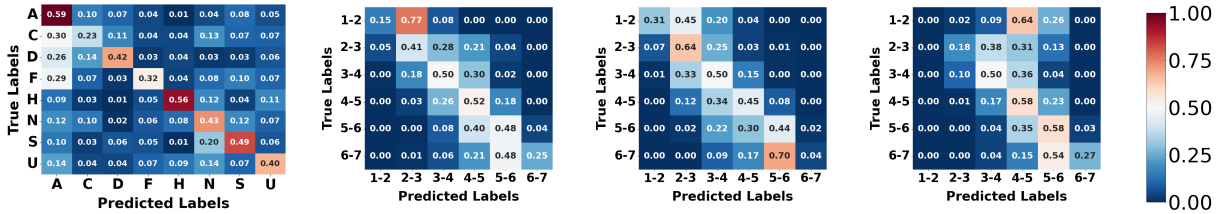
Use of ensemble: Ensemble learning emerged as a key strategy for improving generalization (95%). The NTUA team achieved high performance by employing an ensemble of 14 independently trained fusion models. This strategy highlights the effectiveness of combining multiple foundational models through deep hierarchical fusion. Similarly, UNICAMP trained a five-fold stratified RandomForest meta-model on the logits from multiple architectures for more robust predictions. ABHINAYA employed a majority voting strategy across five classifiers, ensuring a balance between stability and accuracy.

Data augmentation: Data augmentation played a significant role in several top submissions (60%), where several teams employed techniques such as noise addition and reverberation or label augmentation using detailed individual annotations. The SAIL team introduced an augmentation strategy similar to Lotfian et al. [37], where two training audios were mixed with partial overlaps, and the resulting label was an average of the two audio samples.

Training process: The training methodologies adopted by participants reflected diverse strategies for optimization and loss functions. Most teams fine-tuned their speech foundational models, allowing them to adapt to SER-specific tasks effectively. Weighted cross-entropy loss was the most commonly used strategy for addressing class imbalance (70%), with focal loss [38] (18%) employed by some teams to focus on hard-to-classify samples. For emotional attribute prediction, most teams used the one-minus concordance correlation coefficient (1-CCC) loss (87%) to ensure stable and consistent performance across arousal, valence, and dominance.

Class imbalance: In addition to loss functions, top-performing teams use other strategies to deal with class imbalance. The SAIL team introduced label augmentation by removing a subset of annotator labels at random and recalculating the label distribution, improving robustness against noisy annotations. Other teams used batch balancing and inverse frequency weighting to ensure more equitable learning across emotion classes.

Computational efficiency strategies: Adaptive learning rate scheduling, such as cosine annealing and exponential decay, was another common feature among top submissions. The use of gradient accumulation allowed many teams to train larger models with limited hardware resources, optimizing their training processes without compromising performance.



(a) Categorical – NTUA (b) Arousal – SAIL (c) Valence – BIIC (d) Dominance – SAIL

Figure 3: *Confusion matrices of the top-performing models for categorical and attribute-based prediction. (a) Categorical model from NTUA evaluated on eight emotions: A = Anger, C = Contempt, D = Disgust, F = Fear, H = Happiness, N = Neutral, S = Sadness, U = Surprise. (b-d) The best models for arousal, valence, and dominance (SAIL, BIIC, and SAIL, respectively). For attributes, continuous scores were grouped into discrete classes for evaluation. The color bar represents the count scale used in all confusion matrices.*

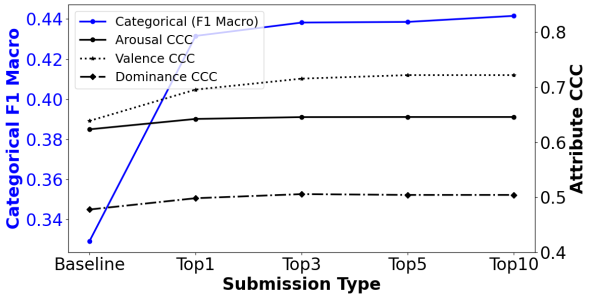


Figure 4: *Performance comparison of categorical and attribute-based emotion recognition tasks. The plot shows baseline performance, top submissions, and combinations of top submissions for both tasks.*

5. Top Performing Model Evaluations

Figure 3(a) presents the confusion matrix for the top-performing submission in categorical emotion recognition (NTUA). Contempt and fear remain the most challenging classes, being frequently misclassified as anger. Although some improvement is observed in recognizing fear compared to previous challenge submissions [7], class imbalance remains a significant issue, as these emotions are underrepresented in the training and development set. Interestingly, among minority emotions, negative emotions (fear, disgust, and contempt) are more prone to being misclassified compared to their positive counterparts (surprise).

Figures 3(b), 3(c) and 3(d) present the confusion matrices for the top-performing submissions for arousal (SAIL), valence (BIIC) and dominance (SAIL), respectively. We create the figures by dichotomizing the continuous predicted and ground truth values into six bins/classes. The confusion matrices for emotional attributes show that predicting extreme attribute scores (e.g., very high or very low values for arousal, valence, and dominance) is particularly challenging. As illustrated, most samples with extreme attribute values tend to be predicted toward the neutral bins. This pattern is especially prominent for low-arousal, low-dominance, and high-valence samples, highlighting the difficulty of accurately capturing these extremes in naturalistic speech. The results emphasize the need for models that can better represent and predict extreme attribute values while maintaining overall stability.

Figure 4 compares the baseline with fusion results of top submissions for categorical and attribute-based tasks. The plot includes top1 submissions and ensemble-based approaches (Top3, Top5, Top10). Categorical labels are derived using a plurality vote, and attribute scores are obtained by averaging predictions across the top submissions. The performance of categorical emotion recognition and valence prediction shows significant improvement through ensemble methods, with valence reaching a CCC of 0.734 when combining the top five submissions. In contrast, arousal and dominance prediction show minimal benefit from ensemble strategies. The primary distinction between top submissions lies in their use of diverse text-based foundational models, indicating that valence prediction can be substantially improved through model diversity. These findings confirm the critical role of ensemble techniques and multimodal approaches in achieving state-of-the-art performance in SER.

6. Conclusions

The Interspeech 2025 Challenge highlighted numerous challenges in advancing SER in naturalistic settings. One of the most persistent issues across submissions was class imbalance, particularly in categorical emotion recognition. Emotions such as contempt and fear remained difficult to classify due to their limited representation in the training and development sets. Despite the use of advanced loss functions such as focal loss and class-weighted cross-entropy, achieving balanced performance across all categories remains a challenge. Additionally, predicting extreme values in emotional attributes such as very high arousal or very low dominance proved difficult, with most models gravitating toward mid-range values. This observation suggests the need for more robust models that can better handle the full spectrum of emotional expressions. Several submissions highlighted the computational challenges of training large multimodal models combining speech and text features. Teams that integrated multiple self-supervised learning models and text-based foundational models faced high computational costs and complex optimization strategies. The trade-off between model complexity and generalization remains a significant hurdle. An open challenge is investigating strategies to mitigate gender bias and variability across speaker demographics, which require more targeted solutions for equitable SER performance across all groups.

7. Acknowledgement

This work was funded by NSF under grant CNS-2016719.

8. References

[1] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. Calgary, AB, Canada: IEEE, April 2018, pp. 5084–5088.

[2] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 415–420.

[3] ——, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.

[4] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1215–1227, April-June 2023.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[6] R. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[7] 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.

[8] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[9] S. Livingstone and F. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, May 2018.

[10] 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, October-December 2019.

[11] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *ArXiv e-prints (arXiv:1909.00360)*, pp. 1–19, May 2019.

[12] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[13] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[14] 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, October 2022.

[15] A. T. Liu, S.-W. Li, and H.-Y. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, July 2021.

[16] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[17] 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.

[18] T. Wolf *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[19] G. Chatzichristodoulou, D. Kosmopoulou, A. Kritikos, A. Poulopoulou, E. Georgiou, A. Katsamanis, V. Katsouros, and A. Potamianos, "Medusa: A multimodal deep fusion multi-stage training framework for speech emotion recognition in naturalistic conditions," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025.

[20] T. Feng, T. Lertpetchpun, D. Byrd, and S. Narayanan, "Developing a top-tier framework in naturalistic conditions challenge for categorized emotion prediction: From speech foundation models and learning objective to data augmentation and engineering choices," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025.

[21] S. Dutta, S. Balaji, V. R, V. Salinamakki, and S. Ganapathy, "ABHINAYA - A System for Speech Emotion Recognition In Naturalistic Conditions Challenge ," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025.

[22] H. J. Jon, L. Jin, H. Jung, H. Kim, D. Min, and E. Y. Kim, "Mater: Multi-level acoustic and textual emotion representation for interpretable speech emotion recognition," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025.

[23] L. Ueda, J. Lima, L. Marques, and P. Costa, "Improving speech emotion recognition through cross modal attention alignment and balanced stacking model," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025.

[24] X. Shi, J. Mi, X. Li, and T. Toda, "Advancing emotion recognition via ensemble learning: Integrating speech, context, and text representations," in *Interspeech 2025*, vol. To appear, Rotterdam, The Netherlands, August 2025.

[25] F. Costa, M. India, and J. Hernando, "Double multi-head attention multimodal system for odyssey 2024 speech emotion recognition challenge," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 266–273.

[26] B. Zgórzyński, J. Wójtowicz-Kruk, P. Masztalski, and W. Średniawa, "Multi-task learning for speech emotion recognition in naturalistic conditions," in *Interspeech 2025*, vol. To appear, Rotterdam, The Netherlands, August 2025.

[27] T. Lertpetchpun, T. Feng, D. Byrd, and S. Narayanan, "Developing a high-performance framework for speech emotion recognition in naturalistic conditions challenge for emotional attribute prediction," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025.

[28] Y. Liu, Y. Gu, J. Luo, W. Zheng, C. Lu, and Y. Zong, "Interactive fusion of multi-view speech embeddings via pretrained large-scale speech models for speech emotional attribute prediction in naturalistic conditions," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025.

[29] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *ArXiv e-prints (arXiv:2104.01027)*, pp. 1–9, April 2021.

[30] 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.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[32] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP*, Nov. 2020, pp. 1644–1650.

[33] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[34] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[35] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," 2024. [Online]. Available: https://arxiv.org/abs/2412.13663

[36] T. Changli, Y. Wenyi, S. Guangzhi, C. Xianzhao, T. Tian, L. Wei, L. Lu, M. Zejun, and Z. Chao, "SALMONN: Towards generic hearing abilities for large language models," *arXiv:2310.13289*, 2023.

[37] R. Lotfian and C. Busso, "Over-sampling emotional speech data based on subjective evaluations provided by multiple individuals," *IEEE Transactions on Affective Computing*, vol. 4, no. 12, pp. 870–882, October-December 2021.

[38] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2980–2988.