Explainable speech emotion recognition through attentive pooling: insights from attention-based temporal localization

Tahitoa Leygue, Astrid Sabourin, Christian Bolzmacher, Sylvain Bouchigny, Margarita Anastassova, Quoc-Cuong Pham

Explainable Speech Emotion Recognition Through Attentive Pooling: Insights from Attention-Based Temporal Localization

Tahitoa Leygue, Astrid Sabourin, Christian Bolzmacher, Sylvain Bouchigny, Margarita Anastassova, Quoc-Cuong Pham

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

tahitoa.leygue@cea.fr, quoc-cuong.pham@cea.fr

Abstract

State-of-the-art transformer models for Speech Emotion Recognition (SER) rely on temporal feature aggregation, yet advanced pooling methods remain underexplored. We systematically benchmark pooling strategies, including Multi-Query Multi-Head Attentive Statistics Pooling, which achieves a 3.5 percentage point macro F1 gain over average pooling. Attention analysis shows 15 percent of frames capture 80 percent of emotion cues, revealing a localized pattern of emotional information. Analysis of high-attention frames reveals that non-linguistic vocalizations and hyperarticulated phonemes are disproportionately prioritized during pooling, mirroring human perceptual strategies. Our findings position attentive pooling as both a performant SER mechanism and a biologically plausible tool for explainable emotion localization. On Interspeech 2025 Speech Emotion Recognition in Naturalistic Conditions Challenge, our approach obtained a macro F1 score of 0.3649.

Index Terms: speaker emotion recognition, affective computing, multimodal learning, multihead attention

1. Introduction

Speech serves as humanity's primary communication medium, conveying both linguistic content and paralinguistic cues about emotions, intentions, and context. The complexities of vocal modulations (pitch, speech rate, variance) provide crucial insights into speakers' emotional states [1]. Emotions—complex psychological and physiological responses to stimuli—are fundamental to human communication and social interaction [2].

Traditional SER systems relied on handcrafted acoustic features—combining prosodic elements (pitch, energy, duration) and spectral descriptors (MFCCs, formants) [3]—and used conventional classifiers such as SVMs [4], HMMs [5], and GMMs [6]. While effective for acted datasets, these approaches underperformed with spontaneous emotions in real-world scenarios [7]. Early integration of basic textual features [8] was also limited in capturing complex semantic relationships.

Deep learning has revolutionized SER with end-to-end architectures that automatically learn features. Early innovations used CNNs for spectrogram analysis [9] and attention-enhanced LSTMs [10], leading to advanced models like Wav2Vec2.0 [11] and HuBERT [12]. Originally for speech recognition, these models now excel in various audio tasks including SER by learning powerful representations from unlabeled data. Recent developments like W2V-BERT 2.0 (600M parameters, trained on 4.5M hours of audio) [13] further underscore their potential for SER tasks.

For textual analysis, BERT-based architectures [14] excel at capturing contextual information, making them a popular choice for tasks that involve processing text derived from audio transcripts. Compared to large language models (LLMs) such as LLaMA [15], BERT offers a compelling trade-off between performance and computational efficiency.

Multimodal approaches which combine audio representations with their corresponding textual transcripts have shown improved robustness [16], though determining optimal strategies for aggregating these features remains an open area of research. The integration of lexical information alongside acoustic features has consistently demonstrated superior performance in emotion recognition tasks compared to unimodal approaches. Recent architectures leverage pre-trained models for both modalities, such as Wav2Vec 2.0 for speech and BERT variants or LLMs for text, thereby benefiting from transfer learning. These approaches have proven particularly effective in the Odyssey 2024 Challenge [17], where multimodal systems consistently outperformed their unimodal counterparts [18, 19].

The transition from frame-level to utterance-level representations is a pivotal step in SER architectures, particularly in the era of self-supervised speech models. Models like Wav2Vec 2.0 have proven adept at extracting rich frame-wise features, but the challenge lies in aggregating these fine-grained temporal representations into effective utterance-level embeddings. While the temporal dynamics of speech are inherently complex, the choice of aggregation mechanisms significantly impacts the performance of downstream emotion recognition tasks.

Historically, simple pooling methods like mean, max, and downsampling gained popularity from text embeddings [20], but struggled with speech's complex temporal dependencies. Statistics Pooling [21] advanced this by incorporating higher-order moments, capturing richer patterns crucial for self-supervised SER.

A significant advancement came with Attentive Statistics (AS) Pooling [22], which introduced learnable parameters to dynamically weight frames, enabling better capture of long-range dependencies. This innovation led to Multi-Head variants including Self-Attentive (SA) pooling [23] and Multi-Head Attention (MHA) Pooling [24]. The most recent advancement, Multi-Query Multi-Head Attentive (MQMHA) Pooling [25], can be viewed as a generalization of these approaches, combining multiple attention mechanisms with statistical aggregation.

Speaker identification determines who is speaking, while SER analyzes emotional content in speech. Both examine vocal characteristics, with MQMHA effectively capturing subtle acoustic features for each. For identification, attention heads focus on consistent traits like timbre; for SER, they track dynamic features such as prosody and speech rate indicating emotions. Both tasks benefit from MQMHA's detailed analysis while targeting different speech characteristics.

We present a multimodal framework for SER that leverages the MSP-Podcast corpus dataset [26]. Developed for the Task 1

on the Interspeech 2025 Speech Emotion Recognition in Naturalistic Conditions Challenge, our framework introduces three key innovations. We first introduce a flexible architecture paired with a training procedure that gradually unfreezes pretrained audio and text encoders. This approach enables the integration of diverse pooling strategies, mitigates overfitting, and ensures computational efficiency. Secondly, we systematically evaluate various pooling mechanisms and adopt MQMHA Pooling, which consistently improves performance on the speech emotion recognition task. Finally, through attention weight analysis, we provide empirical evidence that emotional markers are temporally localized in speech, with attention patterns aligning with specific speech characteristics such as emphasized syllables. These findings advance our understanding of how temporal dynamics contribute to emotion recognition in naturalistic conditions.

2. Methodology

2.1. Model architecture

2.1.1. Audio and Text Encoders

Our architecture leverages state-of-the-art pretrained models for both audio and text processing, is represented in Figure 1. For audio encoding, we employ W2V-BERT 2.0 [13], which has demonstrated superior performance in paralinguistic tasks through its self-supervised pretraining on large-scale speech data. We use the pretrained checkpoint available at `https://huggingface.co/facebook/w2v-bert-2.0`.

For text encoding, we utilize DeBERTa v3 [27], which offers several advantages over traditional BERT-based models for our emotion recognition task. The model's disentangled attention mechanism and enhanced positional encoding improve the capture of nuanced emotional content in transcribed speech by representing each word with two distinct vectors: one for position and one for content. While large language models like LLaMA 3 [15] have shown impressive results in various NLP tasks, we opted for DeBERTa v3 due to its computational efficiency and proven effectiveness in emotion-specific tasks [28]. The pretrained model is available at `https://huggingface.co/microsoft/deberta-v3-base`.

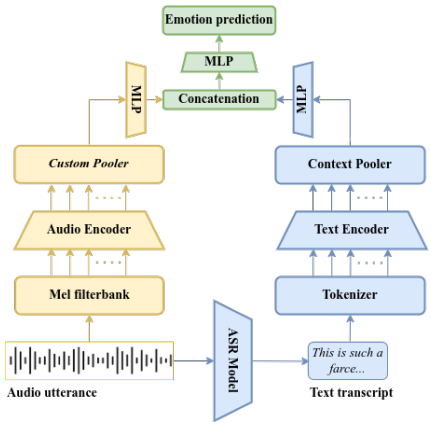

Figure 1: *Multimodal model architecture. "Custom Pooler" denotes one of the pooling strategy described in Section 2.2*

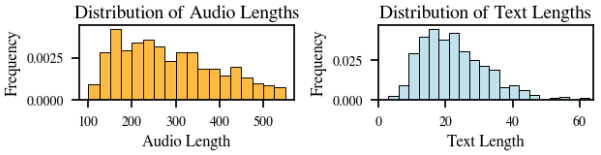

Figure 2: *Distribution of the number of frames per dataset item for audio and text modalities.*

2.1.2. Multimodal Fusion

The fusion module concatenates the pooled embeddings from both modalities, followed by a Multi-Layer Perceptron (MLP) classifier. This approach, while simple, has shown robust performance in multimodal emotion recognition tasks and during Odyssey 2024 challenge [17, 18, 19].

2.2. Pooling Strategies

A key insight from our analysis of the dataset highlights the significant disparity in length between audio data and its text transcriptions. Specifically, audio sequences contain a substantially higher number of frames compared to the token count in their corresponding textual representations, as illustrated by the histograms of data lengths in Figure 2. This stark difference underscores the unique challenges posed by the high temporal resolution of audio data, motivating our investigation into advanced pooling strategies for effectively aggregating these fine-grained frame-level features into robust utterance-level representations.

2.2.1. Basic Pooling Methods

To ensure a transparent basis for evaluating the performance of our custom pooling layers, we implement standard aggregation methods that account for masking to handle variable-length sequences. This masking mechanism ensures robustness by avoiding any influence from padding tokens, thereby eliminating sources of error.

Each method operates on a batch of input feature matrices $\mathbf{X} \in \mathbb{R}^{B\times T\times K}$, where B is the batch size, T the maximum sequence length, and K the feature size, with corresponding masking matrices $\mathbf{M} \in \{0,1\}^{B\times T}$ to denote valid frames.

1. Max Pooling can be expressed as $\mathbf{Y} \in \mathbb{R}^{B\times K}$ where:

$$\mathbf{Y}_b = \max_{t\in\{1,\dots,T\}} \left(\mathbf{X}_{b,t} \cdot M_{b,t}\right) \tag{1}$$

2. Average Pooling can be expressed as $\mathbf{Y} \in \mathbb{R}^{B\times K}$ where:

$$\mathbf{Y}_b = \boldsymbol{\mu}_b = \frac{\sum_{t=1}^{T} \mathbf{X}_{b,t} \cdot M_{b,t}}{\sum_{t=1}^{T} M_{b,t}} \tag{2}$$

3. Statistics Pooling is $\mathbf{Y} \in \mathbb{R}^{B\times 2K}$ where $\mathbf{Y}_b = [\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b]$ and $(\cdot)^{\odot 2}$ denotes the element-wise squaring:

$$\boldsymbol{\sigma}_b = \sqrt{\frac{\sum_{t=1}^{T} M_{b,t} \cdot (\mathbf{X}_{b,t} - \boldsymbol{\mu}_b)^{\odot 2}}{\sum_{t=1}^{T} M_{b,t}}} \tag{3}$$

2.2.2. Advanced Pooling strategies

We implement MQMHA Pooling [25], which generalizes existing attention-based pooling methods. Using the same notations as above, we also introduce $Q \in \mathbb{N}, H \in \mathbb{N}$ respectively the number of queries and heads.

Let $K' = K/H \in \mathbb{N}$, we partition $\mathbf{X}$ into H equal parts such that $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(H)}], \quad \mathbf{X}^{(h)} \in \mathbb{R}^{B \times T \times K'}$. Let $F_{n,p}^{(q,h)}$ denote either a linear mapping (for $n = 1$) or a two-layer MLP with a hidden layer of dimension p and a ReLU activation (for $n = 2$); q and h are the query and head indices, respectively. We sequentially apply the $F_{n,p}^{(q,h)}(\cdot)$ and discard masked layers:

$$e_{b,t}^{(q,h)} = F_{n,p}^{(q)}\left(\mathbf{X}_{b,t}^{(h)}\right) \tag{4}$$

$$\tilde{e}_{b,t}^{(q,h)} = \begin{cases} e_{b,t}^{(q,h)} & \text{if } M_{b,t} = 1 \\ -\infty & \text{otherwise} \end{cases} \tag{5}$$

We finally apply softmax to get the attention weights, with $\omega_{b,t}^{(q,h)}$ being the weight associated to the score $\tilde{e}_{b,t}^{(q,h)}$.

Representation of weighted mean and standard deviation can be computed from the following equations.

$$\boldsymbol{\mu}_b^{(q,h)} = \sum_{t=1}^{T} \omega_{b,t}^{(q,h)} \mathbf{X}_{b,t}^{(h)} \tag{6}$$

$$\boldsymbol{\sigma}_b^{(q,h)} = \sqrt{\sum_{t=1}^{T} \omega_{b,t}^{(q,h)} \left(\mathbf{X}_{b,t}^{(h)}\right)^{\odot 2} - \left(\boldsymbol{\mu}_b^{(q,h)}\right)^{\odot 2}} \tag{7}$$

The output $\mathbf{Y} \in \mathbb{R}^{B \times 2QK}$ is obtained by concatenating the outputs from all queries and heads. This formulation encompasses several existing attention mechanisms as special cases.

2.3. Training protocol

Our training strategy employs gradual unfreezing, a technique that balances the stability-flexibility trade-off in transfer learning [29]. Fully unfreezing all layers initially risks catastrophic forgetting (where over-parameterized models overwrite pretrained knowledge and overfit to sparse emotion labels), while freezing all but the last layers limits adaptation capacity. Gradual unfreezing navigates this compromise by incrementally exposing pretrained parameters to task-specific features, enabling controlled evolution of multimodal representations. This preserves generalizable acoustic-linguistic patterns while adapting to emotional semantics, facilitating robust convergence through curriculum-inspired parameter updates.

Our protocol consists of three phases. First, we freeze both encoders and train only the classification head. This initial phase allows the fusion layer to learn cross-modal relationships while preserving the pretrained representations. Subsequently, we selectively unfreeze the upper transformer layers of both encoders, following the intuition that higher layers capture more task-specific features. Finally, we unfreeze all parameters, including the embedding layers and feature extractors. Transition between steps are triggered by an early stopping module.

To address class imbalance we employ a class-weighted focal loss:

$$\mathcal{L} = -\alpha_c (1 - p_c)^{\gamma} \log(p_c) \tag{8}$$

where p_c is the model's estimated probability for the target class, α_c is the class weight computed as the inverse of class frequency, and $\gamma = 2$.

For optimization, we use AdamW with weight decay $\lambda = 0.01$ and implement a linear learning rate warm-up followed by cosine decay. The learning rates are scaled progressively across the phases: $\eta_1 = 1 \times 10^{-5}$ in phase 1, $\eta_2 = 3 \times 10^{-6}$ in phase 2, and $\eta_3 = 1 \times 10^{-6}$ in phase 3.

3. Experimental setup

In this section, we outline our experimental design, detailing the dataset and evaluation protocol. We evaluate our approach on the challenge dataset—derived from the MSP-Podcast corpus [26] and featuring naturalistic emotional speech—by addressing class imbalance. To mitigate frequency bias while preserving diverse examples to learn emotions, we randomly subsample the majority classes in the training set to a maximum imbalance ratio of 8:1. This choice balances the reduction of frequency bias present in the original training set with a maximum imbalance ratio of 26:1 with the need to preserve sufficient diversity for effective emotion learning. Our final training subset comprises 49,248 utterances, 1,994 speakers, and 78 hours of recording. For robust model selection, we create a balanced development set by sampling an equal number of instances per emotion category following the approach proposed by Härm *et al.* [19], yielding 2608 utterances, 473 speakers, and 5 hours of audio. We used the original test set for final evaluation.

Since test set transcripts are not provided, we generate consistent transcriptions across all splits using Whisper [30]. We used pretrained weights available at `https://huggingface.co/openai/whisper-large-v3`.

4. Results and dicussion

4.1. Multimodal approach

Initial experiments demonstrate the effectiveness of our multimodal approach compared to unimodal baselines. As shown in Table 1, our multimodal architecture achieves a macro F1 score of 0.3559 on the development set, outperforming both the audio-only and transcript-only models. This suggests that audio and textual modalities provide complementary information for emotion recognition, with each modality capturing distinct emotional cues that contribute to better overall performance.

4.2. Comparative pooling analysis

As shown in Table 2, attention-based pooling strategies surpass static pooling (average/statistical) baselines when optimally parameterized, though improper configurations degrade performance due to their architectural complexity. In what is following, we set parameter $p = 256$. Multi-head approaches are consistently configured with $n = 1$, whereas alternative methods employ $n = 2$.

MQMHA with $Q = 2, H = 2$ achieves the highest validation performance, indicating its capacity to disentangle multimodal emotion patterns effectively. Due to time constraints, however, only the AS strategy was submitted for evaluation on the test set, despite MQMHA's superior performance. We attribute this decision to diminishing returns observed at higher query counts ($Q > 2$), where increased complexity risks overfitting without commensurate gains in generalization—an outcome aligned with sparse emotional label distributions.

Table 1: *Macro F1 Scores models on MSP-Podcast corpus.*

Model	Dev Score	Test score
Official baseline (2025)		0.32
W2V-BERT 2.0 (audio only)	0.2921	
DeBERTa V3 (transcript only)	0.3174	
Multimodal baseline	0.3559	

Table 2: *Macro F1 Scores on MSP-Podcast corpus using various pooling strategies.*

Pooling strategy	**Dev score**	**Test score**
Official baseline (2025)		0.3293
Average (*Default*)	0.3559	
Max	0.3418	
Statistics	0.3623	
AS ($Q = 1, H = 1$)	0.3884	0.3649
SA ($Q = 2, H = 1$)	0.3416	
MHA ($Q = 1, H = 2$)	0.3829	
MQMHA ($Q = 2, H = 2$)	**0.3912**	
MQMHA ($Q = 2, H = 4$)	0.3675	
MQMHA ($Q = 4, H = 4$)	0.3514	

4.3. Localization of emotional cues

While attention-driven classification gains suggest that models localize emotionally salient regions, it remains unclear whether such areas holistically capture emotional content. We analyze attention heatmaps (from an AS pooling layer) and their temporal dynamics to scrutinize what the model learns—specifically, whether high-attention frames align with interpretable acoustic or linguistic cues, and how this informs emotion recognition efficacy. Figure 3 visualizes these weights for a sample utterance that was correctly classified as expressing sadness. On the balanced development set, an averaged correlation of $\rho = 0.20 \pm 0.13$ between attention weights and audio energy suggests that the attention mechanism effectively focuses on emotionally salient regions rather than merely reflecting raw acoustic intensity.

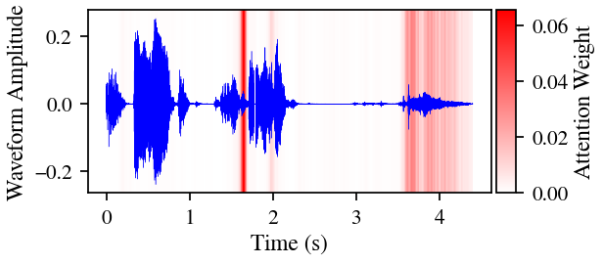

Figure 3: *Attention heatmap on sample 5244_0119.*

Figure 4 presents the aggregated cumulative distribution of attention weights across temporal frames, revealing the model's attention allocation patterns in sequential data. The distribution exhibits a characteristic steep initial slope, where on average, 15% of frames account for 80% of the cumulative attention weight. This highly concentrated distribution suggests that emotional cues are primarily localized within specific temporal regions rather than being uniformly distributed across the utterance. The observed pattern follows a Pareto-like distribution, where a small subset of temporal frames captures the majority of the model's focus. This finding suggests an efficient information extraction mechanism, where the attention layer successfully identifies and emphasizes the most emotionally salient segments.

Figure 5 shows the results of our Bayesian likelihood analysis, which calculated the frequency of attended phonemes relative to their prior corpus frequency. The analysis revealed systematic patterns in the most attention-grabbing phonemes at the utterance level, which were strongly correlated with phonetic prominence and acoustic salience. We leveraged the phoneme annotations available in the MSP-Podcast corpus [26]. The non-linguistic marker `spn` (spoken noise) was exceptionally over-represented, indicating attention weights disproportionately prioritize non-speech vocalizations (e.g., breath sounds, laughter) during pooling. Primary-stressed vowels (`AW1`, `AY1`) and diphthongs (`aw`) exhibited elevated ratios, consistent with their acoustic markedness (longer duration, higher intensity). Secondary stress (`AE2`) also showed heightened salience, corroborating the role of syllabic prominence in perceptual weighting. The attention mechanism's prioritization of hyperarticulated vowels and non-canonical phones aligns with human perceptual strategies for decoding speech in noise [31, 32], reinforcing its biological plausibility in SER systems.

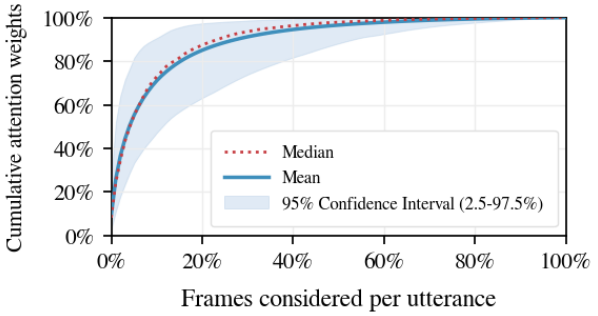

Figure 4: *Attention weight distribution across frames.*

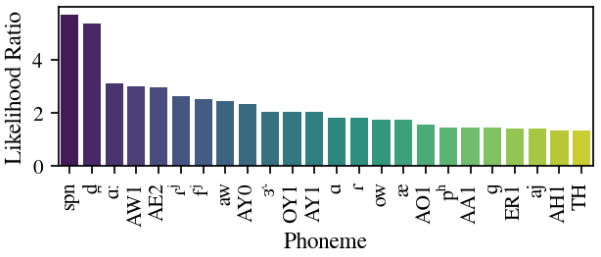

Figure 5: *Attention phoneme salience as bayesian likelihood ratios.*

5. Conclusions

In this work, we introduced MQMHA to Speech Emotion Recognition, demonstrating its advantages over traditional attention mechanisms in the context of the Interspeech 2025 Speech Emotion Recognition Challenge in Naturalistic Conditions Challenge. Our comprehensive evaluation showed that MQMHA outperforms Attentive Statistics and other pooling strategies while maintaining computational efficiency. This advancement is integrated into a multimodal architecture combining W2V-BERT 2.0 and DeBERTa v3 into a lightweight multimodal architecture (less than 1B parameters), supported by considerations of class imbalance and gradual unfreezing strategies. Through attention analysis, we revealed that attention-based pooling methods effectively identifie emotionally salient regions independently of signal energy, contributing to both performance improvement and interpretability in SER systems.

6. Acknowledgements

This publication was made possible by the use of the CEA-List FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council; and has been performed in the scope of the OASEES Project, supported by the Commission of the European Communities */HORIZON, GA No.101092702*.

7. References

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020.

[3] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. Montreal, Que., Canada: IEEE, 2004, pp. I–577–80.

[4] Y. Chavhan, ML. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.

[5] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[6] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

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

[8] Chul Min Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[9] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, Mar. 2016, pp. 5200–5204.

[10] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 2227–2231.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[12] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[13] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[14] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1. Minneapolis, Minnesota, 2019.

[15] 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," 2024.

[16] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning Alignment for Multimodal Emotion Recognition from Speech," 2019.

[17] 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)*. ISCA, Jun. 2024, pp. 247–254.

[18] M. Chen, H. Zhang, Y. Li, J. Luo, W. Wu, Z. Ma, P. Bell, C. Lai, J. Reiss, L. Wang, P. C. Woodland, X. Chen, H. Phan, and T. Hain, "1st Place Solution to Odyssey Emotion Recognition Challenge Task1: Tackling Class Imbalance Problem," May 2024.

[19] H. Härm and T. Alumäe, "TalTech Systems for the Odyssey 2024 Emotion Recognition Challenge," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*. ISCA, Jun. 2024, pp. 255–259.

[20] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-Supervised Speech Representation Learning: A Review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022.

[21] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 999–1003.

[22] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech 2018*, ser. Interspeech_2018. ISCA, Sep. 2018.

[23] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, vol. 2018, 2018, pp. 3573–3577.

[24] M. India, P. Safari, and J. Hernando, "Self Multi-Head Attention for Speaker Recognition," Jul. 2019.

[25] M. Zhao, Y. Ma, Y. Ding, Y. Zheng, M. Liu, and M. Xu, "Multi-Query Multi-Head Attention Pooling and Inter-Topk Penalty for Speaker Verification," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 6737–6741.

[26] 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, Oct. 2019.

[27] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," Mar. 2023.

[28] M. M. Imran, "Emotion Classification In Software Engineering Texts: A Comparative Analysis of Pre-trained Transformers Language Models," in *Proceedings of the Third ACM/IEEE International Workshop on NL-based Software Engineering*. Lisbon Portugal: ACM, Apr. 2024, pp. 73–80.

[29] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," May 2018.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022.

[31] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, first edition ed. Hoboken, N.J: Wiley, 2014.

[32] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, 2003.