

Improving Multimodal Emotion Recognition by Leveraging Acoustic Adaptation and Visual Alignment

Zhixian Zhao

Northwestern Polytechnical University
Xi'an, China
zxzhao@mail.nwpu.edu.cn

Haifeng Chen

Shaanxi University of Science and
Technology
Xi'an, China
chenhaifeng@sust.edu.cn

Xi Li

Shaanxi University of Science and
Technology
Xi'an, China
lixii@sust.edu.cn

Dongmei Jiang

Peng Cheng Laboratory

Shenzhen, China

Northwestern Polytechnical University
Xi'an, China
jiangdm@nwpu.edu.cn

Lei Xie*

Northwestern Polytechnical University
Xi'an, China
lxie@nwpu.edu.cn

Abstract

Multimodal Emotion Recognition (MER) aims to automatically identify and understand human emotional states by integrating information from various modalities. However, the scarcity of annotated multimodal data significantly hinders the advancement of this research field. This paper presents our solution for the MER-SEMI sub-challenge of MER 2024. First, to better adapt acoustic modality features for the MER task, we experimentally evaluate the contributions of different layers of the pre-trained speech model HuBERT in emotion recognition. Based on these observations, we perform Parameter-Efficient Fine-Tuning (PEFT) on the layers identified as most effective for emotion recognition tasks, thereby achieving optimal adaptation for emotion recognition with a minimal number of learnable parameters. Second, leveraging the strengths of the acoustic modality, we propose a feature alignment pre-training method. This approach uses large-scale unlabeled data to train a visual encoder, thereby promoting the semantic alignment of visual features within the acoustic feature space. Finally, using the adapted acoustic features, aligned visual features, and lexical features, we employ an attention mechanism for feature fusion. On the MER2024-SEMI test set, the proposed method achieves a weighted F1 score of 88.90%, ranking fourth among all participating teams, validating the effectiveness of our approach.

CCS Concepts

- Computing methodologies → Artificial intelligence; • Human-centered computing → Human computer interaction (HCI).

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MRAC '24, November 1, 2024, Melbourne, VIC, Australia.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1203-6/24/11
<https://doi.org/10.1145/3689092.3689407>

Keywords

Multimodal Emotion Recognition, Fine-tuning, Contrastive Learning

ACM Reference Format:

Zhixian Zhao, Haifeng Chen, Xi Li, Dongmei Jiang, and Lei Xie. 2024. Improving Multimodal Emotion Recognition by Leveraging Acoustic Adaptation and Visual Alignment. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing (MRAC '24), November 1, 2024, Melbourne, VIC, Australia*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3689092.3689407>

1 Introduction

Automatic emotion recognition is crucial for human-computer interaction (HCI), allowing computers to detect and respond to users' emotional states [4, 12]. Multimodal emotion recognition (MER), based on supervised learning, has shown promising results [6, 10, 20]. However, the annotation process is costly and time-consuming, limiting the scale of labeled MER datasets and hindering performance. The MER-SEMI challenge, a sub-challenge of the Multimodal Emotion Recognition Challenge [13, 14], addresses this issue by providing a labeled emotional dataset alongside substantial unlabeled data, enabling participants to explore effective unsupervised or semi-supervised learning strategies.

Pre-trained transformer models [1, 2, 9, 18] have achieved notable success across various speech tasks, excelling in capturing phonetic structures, temporal dependencies, and acoustic features. Studies [11, 19] suggest that essential task-specific information may reside in the hidden representations of different transformer layers, yet there is limited exploration of their contributions to speech emotion recognition. Typically, current parameter fine-tuning methods [7, 8, 22] involve modifying the entire model, which overlooks the varying significance of features across layers. To improve the performance of pre-trained speech models in emotion recognition, it is beneficial to incorporate adapters when fine-tuning certain intermediate layers. This approach aligns features with emotion recognition requirements, reduces training parameters, and preserves the model's generalization capability.

The visual modality provides rich non-verbal information, such as facial expressions, body language, and gestures, essential for

computer vision and natural language processing tasks [21, 25]. To fully leverage multimodal information, comparative cross-modal pretraining methods [5, 16, 23] have been developed. For instance, CLIP [17] achieves a shared semantic space for visual and textual understanding by jointly training image and text encoders. Baseline results [14] indicate that the visual modality's performance in emotion recognition is relatively weaker compared to the acoustic modality. Therefore, after obtaining optimal acoustic features, the visual modality can be aligned with the acoustic feature space. Contrastive learning methods can establish relationships between acoustic and visual modalities, enhancing the visual modality's ability to capture emotional information more accurately.

Based on the above discussions, we propose a semi-supervised multimodal emotion recognition method comprising three stages: acoustic feature adaptation, visual feature alignment, and multimodal feature fusion. The main flow is shown in Figure 1. First, to optimize acoustic modality features for emotion recognition, we conducted an empirical study exploring the performance of different layers of the HuBERT-large model [9] and the effectiveness of multi-layer feature fusion. Guided by these empirical findings, we propose a simple and effective parameter-efficient fine-tuning method. This method enhances recognition performance by incorporating adapters into well-performing intermediate transformer layers and dynamically fusing hidden representations across these layers using learnable weights. Second, to enhance the emotional representation capability of visual modality features, we perform contrastive learning between the fine-tuned acoustic features and visual features processed by a multilayer perceptron (MLP). We leverage a substantial amount of unlabeled visual and audio data to pretrain the vision MLP in an unsupervised manner, ensuring the visual features adapt to the acoustic feature space. Finally, we employ an attention-based feature fusion module to integrate the acoustic, visual, and textual features, achieving a weighted F1 score of 88.90% on the test set.

2 Method

2.1 Acoustic Feature Adaptation

Based on the findings of the baseline study [14], we select the HuBERT-large model [9], which has demonstrated superior performance in emotion recognition tasks, as the feature extractor. Since pre-trained transformer models (e.g., HubERT and wav2vec2.0 [2]) capture unique hidden representations across different layers during audio processing and that these layers may contain complementary information, we utilize the temporal pooling features ($f_a^i \in \mathbb{R}^{d_a}, i \in 1, 2, \dots, k$) from k consecutive middle layers of the HuBERT-large model for the emotion recognition task.

To adapt these features more effectively to the emotion recognition task while maintaining the generalization capability of the pre-trained model, we introduce adapters in these k transformer layers for efficient parameter fine-tuning. The implementation details of the adapters are depicted in Figure 1(Stage 1). Each adapter consists of a bottleneck structure, including a dimension reduction projection layer that reduces the hidden dimension from d_a to \hat{d} , followed by a ReLU non-linear activation function, and then an up-projection layer that restores the dimension back to d_a . Given an input feature x , the output y of the adapter can be represented

as:

$$y = x + \text{ReLU}(W_{up}\text{ReLU}(W_{down}x + b_{down}) + b_{up}), \quad (1)$$

where $W_{down} \in \mathbb{R}^{d_a \times \hat{d}}$, $W_{up} \in \mathbb{R}^{\hat{d} \times d_a}$, $b_{down} \in \mathbb{R}^{\hat{d}}$, $b_{up} \in \mathbb{R}^{d_a}$ are trainable parameters.

To account for the varying contributions of different layers to the emotion recognition task, we introduce learnable weights ($w_i, i \in \{1, 2, \dots, k\}$) to fuse the output features of these k transformer layers, resulting in the fused feature $f_a^{fusion} = \sum_{i=1}^k w_i \cdot f_a^i$, where $f_a^{fusion} \in \mathbb{R}^{d_a}$. We use two types of loss functions to optimize the model. The first is an unsupervised masked reconstruction loss, inspired by [9], this approach predicts the masked portions of the acoustic features using contextual information to learn more robust acoustic representations. We generate masked features $f_a^{masked} \in \mathbb{R}^{d_a}$ through a multi-layer perceptron (MLP) consisting of two fully connected layers and a ReLU layer, and measure the difference between the original and masked features using the mean squared error (MSE) loss function. The second loss function is a supervised cross-entropy (CE) loss. After passing the masked features through a fully connected layer to obtain the predicted results x^{pred} , we use the CE loss function to calculate the loss between the predicted results and the labels y^{label} . These two loss functions can be expressed as:

$$\mathcal{L}_{ce} = CE(x^{pred}, y^{label}), \quad (2)$$

$$\mathcal{L}_{mlm} = MSE(f_a^{masked}, f_a^{fusion}), \quad (3)$$

Thus, the overall objective function for fine-tuning the model is defined as:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{mlm}. \quad (4)$$

2.2 Visual Feature Alignment

Due to the semantic disparity between the visual features extracted by CLIP-large [17] and the audio features, we perform pre-training feature alignment for the visual modality prior to multimodal feature fusion, as illustrated in Figure 1(Stage 2). Inspired by [15], we first pre-train a vision MLP to align visual representations with the semantic space of pre-trained speech emotion representations. During this process, the weights of the CLIP-large model and the fine-tuned HuBERT-large model are kept frozen, and only the weights of the vision MLP are trained.

We introduce an video-audio contrastive learning method. An input video I is encoded by the CLIP-large model and then average-pooled to obtain $f_v \in \mathbb{R}^{d_v}$. Then, an MLP maps the visual embedding to the same dimensionality as the audio embedding, resulting in mapped visual embedding $f_v \in \mathbb{R}^{d_a}$. This transformation can be represented as:

$$f_v = \text{ReLU}(W_1\text{ReLU}(W_2f_v + b_1) + b_2), \quad (5)$$

where $W_1 \in \mathbb{R}^{d_v \times d_a}$ and $W_2 \in \mathbb{R}^{d_a \times d_a}$, $b_1 \in \mathbb{R}^{d_a}$ and $b_2 \in \mathbb{R}^{d_a}$ are trainable parameters of the vision MLP. The goal of the MLP is to align the visual and audio features so they can operate within a unified feature space.

To compute the similarity between the visual and audio embeddings, we first use the fine-tuned HuBERT-large model to convert the input audio A into an embedding sequence $f_a \in \mathbb{R}^{d_a}$. The video-audio similarity can be defined as: $s(I, A) = f_v^\top f_a / \|f_v\| \cdot \|f_a\|$ and the audio-video similarity as: $s(A, I) = f_v^\top f_a / \|f_a\| \cdot \|f_v\|$. For each

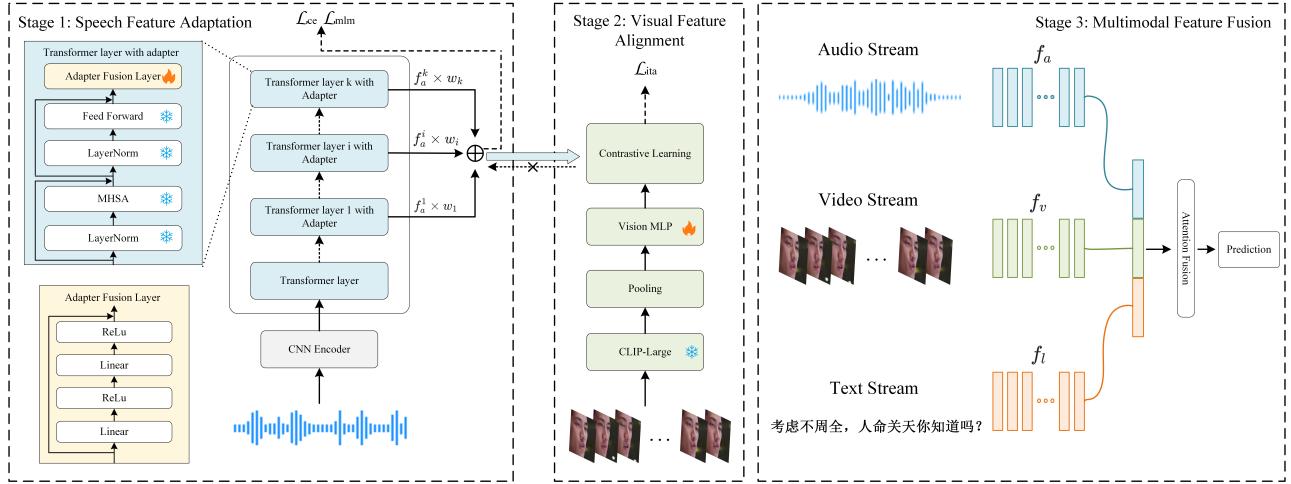


Figure 1: The proposed multimodal emotion recognition model framework

video and audio pair, we calculate the normalized softmax similarity scores to measure the similarities from video to audio and from audio to video:

$$p_j^{i2a}(I) = \frac{\exp(s(I_j, A)/\tau)}{\sum_{j=1}^J \exp(s(I_j, A)/\tau)}, \quad p_j^{a2i}(A) = \frac{\exp(s(A_j, I)/\tau)}{\sum_{j=1}^J \exp(s(A_j, I)/\tau)}, \quad (6)$$

where τ is a learnable temperature parameter. Let y^{i2a} and y^{a2i} represent the ground truth one-hot encoded similarities, with a probability of 0 for mismatched video-audio pairs and 1 for matched pairs. The video-audio contrastive loss \mathcal{L}_{ita} is defined as the cross-entropy H between p and y :

$$\mathcal{L}_{ita} = \frac{1}{2} \mathbb{E}_{(I,A) \sim D} \left[H(y^{i2a}(I), p^{i2a}(I)) + H(y^{a2i}(A), p^{a2i}(A)) \right]. \quad (7)$$

2.3 Multimodal Feature Fusion

Before conducting feature fusion, we outline the extraction processes for the three modalities. Acoustic Features: we utilize the fine-tuned Chinese-HuBERT-Large model [9] to extract the speech representation fa for each audio sample. Visual Features: we input the facial images extracted using the OpenFace toolkit [3] into the CLIP-large model [17] to extract the visual features. Subsequently, we use the vision MLP introduced in Section 2.2 as the feature extractor to obtain the feature representation fv . Lexical Features: we use the Baichuan2 [24] model to extract the feature representation fl from the checked transcripts of the Train&Val set, as well as the subtitle files of the unlabeled data.

After obtaining the feature vectors for the three modalities $fm \in \mathbb{R}^{dm}, m \in (a, v, l)$, we use an MLP composed of several fully connected layers and ReLU activation functions to map each modality's features to the same dimension. Considering the varying importance of each modality, we stack the three modality features together and calculate the attention score α for each modality:

$$h = \text{Concat}(h_a, h_l, h_v), \quad (8)$$

$$\alpha = \text{softmax}(h^\top W_\alpha + b_\alpha), \quad (9)$$

where W_α and b_α are trainable parameters. The final fused feature is given by $z = \alpha h$.

Table 1: Statistics of the MER 2024 dataset

| Dataset | Labeled | Unlabeled | Duration (hr:min:sec) |
|-----------|---------|-------------|-----------------------|
| Train&Val | 5030 | 0 | 05:56:39 |
| MER-SEMI | 0 | 1169/115595 | 100:38:49 |
| MER-NOISE | 0 | 1170/115595 | 100:38:49 |

3 Experiments and Results

3.1 Dataset and Implementation Details

Dataset: We conduct experiments using the MER-SEMI dataset, as detailed in Table 1. The Train&Val set consists of 5,030 video clips with both discrete and dimensional emotion labels. Given the absence of a predefined training/validation split, we utilize five-fold cross-validation on the Train&Val set [14]. To evaluate model generalization, the MER-SEMI track includes 1,169 unlabeled video clips in a test set, drawn from a total of 115,595 unlabeled data. Participants must predict discrete emotion labels across all unlabeled data, not just the test set. The set of discrete emotion labels includes six categories: *neutral*, *anger*, *happiness*, *sadness*, *worry*, and *surprise*. We use a weighted average F1 score as the evaluation metric, aligning with the official baseline.

Implementation Details: For acoustic features adaptation, we select the 16th to 21st layers (total $k = 6$ layers) of the HuBERT-large model and incorporate adapters into these layers for fine-tuning (refer to Section 3.2.1). The feature dimension d within the Adapter is set to 128. During fused feature computation, we assign an initial weight of 1.0 to the 18th layer for its optimal performance, while the other layers are initially weighted at 0.0. Pre-training is conducted on the Train&Val set with a batch size of 16 and a learning rate of 1e-4. We use the Adam optimizer with a weight decay of 0.02. For visual feature alignment, the vision MLP is trained on unlabeled data with a batch size of 1024 and a learning rate of 1e-4. For the feature fusion module, features from the three modalities are mapped to 256 dimensions through an MLP for fusion, with a learning rate set to 1e-4. The dimensions of the visual, audio, and lexical features are $d_v = 768$, $d_a = 1024$, and $d_l = 5120$, respectively.

3.2 Result and Discussion

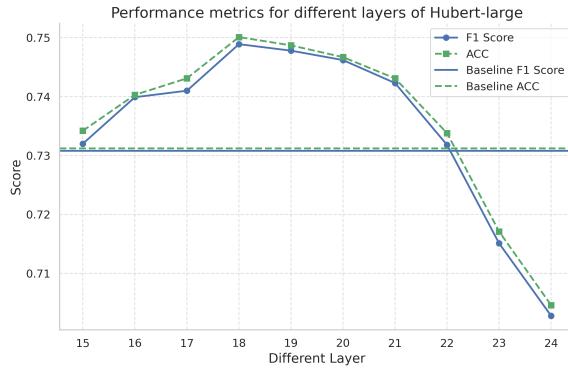


Figure 2: Comparison of the performance of features from different layers of HuBERT-large.

Table 2: Evaluation of Unimodal and Multimodal Feature Performance.

| A | Features | | Train&Val | MER-SEMI |
|--------------------|----------|-----------|------------------------|------------------------|
| | V | T | F1-Score(\uparrow) | F1-Score(\uparrow) |
| Unimodal Results | | | | |
| HL | - | - | 72.82 \pm 0.30 | 83.49 |
| HL(18) | - | - | 74.40 \pm 0.46 | 84.78 |
| HL(16-21) | - | - | 73.98 \pm 0.43 | 84.79 |
| HLFT(18) | - | - | 76.30 \pm 0.33 | 84.69 |
| HLFT(16-21) | - | - | 80.24 \pm 0.21 | 84.88 |
| - | CLIP-L | - | 66.22 \pm 0.43 | 60.95 |
| - | CLIP-L-A | - | 66.05 \pm 0.37 | 64.59 |
| Multimodal Results | | | | |
| HLFT(16-21) | CLIP-L | - | 84.34 \pm 0.25 | 86.78 |
| HLFT(16-21) | CLIP-L-A | - | 84.70 \pm 0.19 | 87.01 |
| HLFT(16-21) | - | Baichuan2 | 79.66 \pm 0.13 | 85.78 |
| HLFT(16-21) | CLIP-L-A | Baichuan2 | 83.85 \pm 0.35 | 88.90 |

3.2.1 Empirical study. To validate the performance of features from different transformer layers in emotion recognition tasks, we extract the features from the last 10 layers of the HuBERT-large model and evaluate their performance on the Train&Val set, as shown in Figure 2. The rationale for focusing on the last 10 layers is that shallow layers typically encode low-level features, such as pitch and short-term energy fluctuations, while deeper layers are more adept at capturing high-level features and global semantic information, which are crucial for emotion recognition. As shown in the figure, features from the 18th layer demonstrate the best performance, significantly surpassing the baseline score, indicating that this layer captures more representative emotional features. Furthermore, features from the 16th through 21st layers consistently outperform the baseline, indicating that these intermediate layers are more suitable for emotion recognition tasks. These layers balance capturing both low-level and high-level information, providing rich audio content while minimizing noise interference. This finding is of significant importance for subsequent model fine-tuning efforts.

3.2.2 Unimodal Recognition Results. We present the experimental results for both unimodal and multimodal approaches in Table 2. For the acoustic modality, HL represents features extracted using the baseline method [14] from the HuBERT-large model, with HL(i) indicating the feature output from the i -th layer. HLFT(i) indicates the

output from the i -th layer after fine-tuning the HuBERT-large model with adapters. As shown in the table, the performance of the fine-tuned features demonstrates a notable improvement over the baseline model. The proposed parameter-efficient fine-tuning method achieves superior performance with multi-layer fused features compared to single-layer features. Specifically, HL(16-21) surpasses HL(18), and HLFT(16-21) outperforms HLFT(18), suggesting that complementary information exists among features from different layers, resulting in more robust results. Moreover, the multi-layer fused features obtained through the proposed parameter-efficient fine-tuning method achieve the highest F1 score of 84.88% on the test set. This method also demonstrates performance improvements of 7.42% on the Train&Val set and 1.39% on the test set compared to the baseline model, validating its effectiveness.

For the visual modality, CLIPL represents the features extracted using the CLIP-Large model, whereas CLIPL-A denotes the features aligned through the proposed visual feature alignment strategy. As shown in Table 2, in comparison to the CLIPL features, the CLIPL-A features exhibit comparable performance on the Train&Val set and show a 3.64% improvement on the test set. These results affirm the efficacy of the visual feature alignment strategy in enhancing performance within multimodal emotion recognition tasks.

3.2.3 Multimodal Recognition Results. We conduct a comprehensive comparison of the multimodal fusion effects, with specific results presented in Table 2. Initially, we fuse the best-performing acoustic modality features, HLFT (16-21), with the visual modality features extracted by CLIP-large. This fusion improves the recognition accuracy from 84.88% to 86.78%. Furthermore, using the aligned features extracted by the pre-trained vision MLP, the recognition accuracy further increases to 87.01%, which provides additional validation for the effectiveness of the feature alignment pre-training method. Similarly, we evaluate the fusion of lexical modality and acoustic modality. When fusing Baichuan2 features with HLFT (16-21) features, the performance on the test set reaches 85.78%. Ultimately, the fusion of all three modality features results in the highest performance of 88.90% on the test set. Additionally, the table illustrates that the effect of multimodal fusion on the test set surpasses that of any single modality.

4 Conclusion

In this study, we propose a multimodal emotion recognition framework for the MER2024-SEMI challenge. Initially, our focus is on fully leveraging acoustic modality features to enhance emotion recognition tasks. We evaluate the performance of different transformer layers of the HuBERT-large model in speech emotion recognition and employ an PEFT method to fine-tune the HuBERT-large model. Subsequently, to enhance the emotional representation of the visual modality, we introduce an unsupervised feature alignment scheme that employs contrastive learning to align visual embeddings with acoustic embeddings. Experimental results validate the effectiveness of the proposed methods, with our approach securing fourth place in the MER2024-SEMI sub-challenge.

5 Acknowledgments

This work is supported by the National Natural Science Foundation of China (grant 62236006, grant 62306172), the Key Research and Development Program of Shaanxi (No. 2022ZDLGY06-03).

References

- [1] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. 2023. Efficient Self-supervised Learning with Contextualized Target Representations for Vision, Speech and Language. In *Proceedings of the 40th International Conference on Machine Learning*. 1416–1429.
- [2] Alexei Baevski, Yuhan Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, Vol. 33. 12449–12460.
- [3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 59–66. <https://doi.org/10.1109/FG.2018.00019>
- [4] Felipe Zago Canal, Tobias Rossi Müller, Jheniffer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sôbieranski. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences* 582 (2022), 593–617. <https://doi.org/10.1016/j.ins.2021.10.005>
- [5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP Learning Audio Concepts from Natural Language Supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10095889>
- [6] Ruijia Fan, Hong Liu, Yidi Li, Peini Guo, Guoquan Wang, and Ti Wang. 2024. AttANet: Attention Aggregation Network for Audio-Visual Emotion Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8030–8034. <https://doi.org/10.1109/ICASSP48485.2024.10447640>
- [7] Yuan Gao, Hao Shi, Chenhui Chu, and Tatsuya Kawahara. 2024. Enhancing Two-Stage Finetuning for Speech Emotion Recognition Using Adapters. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 11316–11320. <https://doi.org/10.1109/ICASSP48485.2024.10446645>
- [8] Erik Goron, Lena Asai, Elias Rut, and Martin Dinov. 2024. Improving Domain Generalization in Speech Emotion Recognition with Whisper. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 11631–11635. <https://doi.org/10.1109/ICASSP48485.2024.10446697>
- [9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460. <https://doi.org/10.1109/TASLP.2021.3122291>
- [10] Qi Huang, Pingting Cai, Tanyue Nie, and Jinshan Zeng. 2024. CLIP-MSA: Incorporating Inter-Modal Dynamics and Common Knowledge to Multimodal Sentiment Analysis With Clip. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8145–8149. <https://doi.org/10.1109/ICASSP48485.2024.10446825>
- [11] Pratik Kumar, Vrunda N. Sukhadia, and S. Umesh. 2022. Investigation of Robustness of Hubert Features from Different Layers to Domain, Accent and Language Variations. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6887–6891. <https://doi.org/10.1109/ICASSP43922.2022.9746250>
- [12] Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. 2022. EEG Based Emotion Recognition: A Tutorial and Review. *ACM Comput. Surv.* 55, 4, Article 79 (nov 2022), 57 pages. <https://doi.org/10.1145/3524499>
- [13] Zheng Lian, Haiyang Sun, Licali Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jimming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2023. MER 2023: Multi-label Learning, Modality Robustness, and Semi-Supervised Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. 9610–9614. <https://doi.org/10.1145/3581783.3612836>
- [14] Zheng Lian, Haiyang Sun, Licali Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jimming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2024. MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary Multimodal Emotion Recognition. <https://arxiv.org/abs/2404.17113>
- [15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, Vol. 36. 34892–34916.
- [16] Yu Pan, Yanni Hu, Yuguang Yang, Wen Fei, Jinxun Yao, Heng Lu, Lei Ma, and Jianjun Zhao. 2024. GEmo-CLAP: Gender-Attribute-Enhanced Contrastive Language-Audio Pretraining for Accurate Speech Emotion Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 10021–10025. <https://doi.org/10.1109/ICASSP48485.2024.10448394>
- [17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. PMLR, 8748–8763.
- [18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202. PMLR, 28492–28518.
- [19] Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2023. On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language* 77 (2023), 101429. <https://doi.org/10.1016/j.csl.2022.101429>
- [20] Xin Sun, Xiangyu Ren, and Xiaohao Xie. 2024. A Novel Multimodal Sentiment Analysis Model Based on Gated Fusion and Multi-Task Learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8336–8340. <https://doi.org/10.1109/ICASSP48485.2024.10446040>
- [21] Juan Terven, Diana-Margarita Córdoba-Esparza, and Julio-Alejandro Romero-González. 2023. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction* 5, 4 (2023), 1680–1716. <https://doi.org/10.3390/make5040083>
- [22] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2022. A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding. <https://arxiv.org/abs/2111.02735>
- [23] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10095969>
- [24] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, and Chenxu Lv. 2023. Baichuan 2: Open Large-scale Language Models. <https://arxiv.org/abs/2309.10305>
- [25] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 8 (2024), 5625–5644. <https://doi.org/10.1109/TPAMI.2024.3369699>