Hierarchical Audio-Visual Information Fusion with Multi-label Joint Decoding for MER 2023

Haotian Wang
Yuxuan Xi
Hang Chen
Jun Du*
University of Science and Technology of China
Hefei, Anhui, China

Yan Song
Qing Wang
Hengshun Zhou
Chenxi Wang
University of Science and Technology of China
Hefei, Anhui, China

Jiefeng Ma
Pengfei Hu
Ya Jiang
Shi Cheng
University of Science and Technology of China
Hefei, Anhui, China

Jie Zhang
University of Science and Technology of China
Hefei, Anhui, China

Yuzhe Weng
Northwestern Polytechnical University
Xi'an, Shaanxi, China

ABSTRACT

In this paper, we propose a novel framework for recognizing both discrete and dimensional emotions. In our framework, deep features extracted from foundation models are used as robust acoustic and visual representations of raw video. Three different structures based on attention-guided feature gathering (AFG) are designed for deep feature fusion. Then, we introduce a joint decoding structure for emotion classification and valence regression in the decoding stage. A multi-task loss based on uncertainty is also designed to optimize the whole process. Finally, by combining three different structures on the posterior probability level, we obtain the final predictions of discrete and dimensional emotions. When tested on the dataset of multimodal emotion recognition challenge (MER 2023), the proposed framework yields consistent improvements in both emotion classification and valence regression. Our final system achieves state-of-the-art performance and ranks third on the leaderboard on MER-MULTI sub-challenge.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; **Artificial intelligence**; **Multi-task learning**; *Computer vision.*

KEYWORDS

MER2023, deep feature fusion, joint decoding, multi-task learning

ACM Reference Format:
Haotian Wang, Yuxuan Xi, Hang Chen, Jun Du, Yan Song, Qing Wang, Hengshun Zhou, Chenxi Wang, Jiefeng Ma, Pengfei Hu, Ya Jiang, Shi Cheng, Jie Zhang, and Yuzhe Weng. 2023. Hierarchical Audio-Visual Information Fusion with Multi-label Joint Decoding for MER 2023. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581783.3612859

*corresponding author

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612859

1 INTRODUCTION

Multimodal emotion recognition (MER) plays a crucial role in natural human-machine interaction [25, 29], intelligent education tutoring [23, 24], and mental health diagnoses [16, 31], etc. In our daily life, when we engage in dialogue and communication, we usually convey our emotions through both verbal and non-verbal content, such as facial expressions and body language [26]. Previous studies focused on emotion recognition in text [12, 33], facial expression [26, 35] and audio [17, 27]. However, it has been observed that research on single-modality approaches has reached a certain bottleneck, thereby leading to increased attention toward the use of multimodal approaches [3, 9, 14, 18].

Regarding human communication scenarios, emotions are mainly expressed through speech and facial expressions, each providing complementary information. As a result, researchers are dedicated to fusing audio and video modal features [10, 22, 36, 37]. For example, Han *et al.* [10] proposed a hierarchical approach that maximized the Mutual Information (MI) among unimodal inputs. Hazarika *et al.* [22] projected each modality to modality-invariant and modality-specific spaces to learn effective representations. Recently, inspired by the success of pre-trained deep features like wav2vec2.0 [1] and HUBURT [13] in other speech-related tasks, some researchers investigate their superiority over hand-engineered features in speech emotion recognition and discovered that these deep features capture more robust acoustic representations [4, 15, 30, 32].

According to the theory of psychological research, there are two main emotional calculation models: discrete theory and dimensional theory. Discrete theory [7] describes emotional states as discrete labels such as "sad" and "happy". On the other hand, the theory of dimensionality [8] suggests that emotional states exist as points in a continuous space. This allows for simulating subtle, complex, and sustained emotional behaviors. Previous studies have found that there exists a strong correlation between these two models [5, 30].

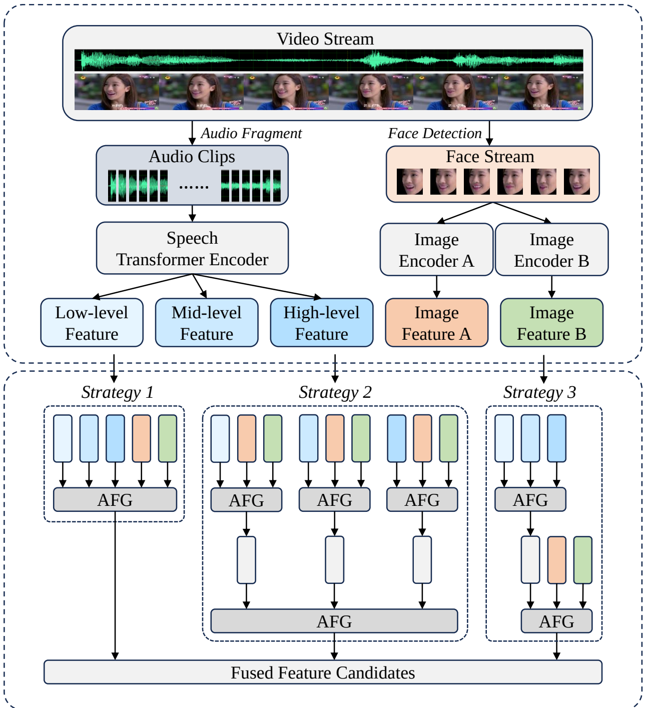

Figure 1: Deep acoustic and visual features extraction and the following three feature fusion structures. Structure of the attention-guided feature gathering (AFG) module is shown in the following figure.

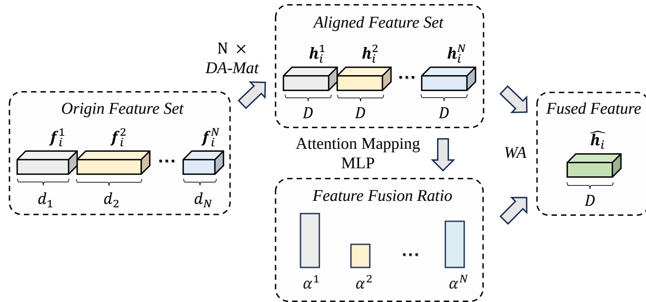

Figure 2: Attention-guided feature gathering (AFG) module. DA-Mat is short for Dimension Align Matrix. WA is short for Weight Average.

In this paper, we propose an efficient multimodal emotion recognition system to recognize both discrete emotion (emotion) and dimensional emotion (valence). Firstly, we extract deep features through different layers of various pre-trained models as robust acoustic and visual representations of raw video segments. Then we fuse these features through three proposed feature fusion structures based on AFG [20]. Afterwards, we have designed a joint decoding module that considers both discrete and dimensional theories based on the correlation between emotion and valence to generate decisions for these two dimensions. A multi-task loss function based on uncertainty [6] is also designed to optimize the whole encoding and decoding process. Finally, we generate final predictions of both discrete and dimensional emotions by decision-level fusion. We evaluate the proposed framework on the dataset of the multimodal emotion recognition challenge (MER 2023) [19]. Experiments show that our framework yields consistent improvements on both emotion classification and valence regression and ranks third on the leaderboard on MER-MULTI sub-challenge.

2 METHODS

In this section, we will discuss our proposed multimodal emotion recognition system in two subsections. The feature extraction and fusion strategies will be illustrated in the first subsection. The proposed joint decoding module of discrete and dimensional emotions and the designed multi-task loss function will be introduced in the second subsection.

2.1 Features encoding and attention-guided fusion

In our proposed architecture, we first extract deep features from pre-trained models as robust acoustic and visual representations of raw video segments. The details are illustrated in Figure 1. Previous research has indicated that different layers in pre-trained speech model HUBURT [13] capture audio hidden states with distinctive characteristics [34]. The hidden states captured by layers closer to the front exhibit increased sensitivity to the acoustic features of the original audio, such as tone or frequency. In contrast, the hidden states captured by layers towards the back demonstrate a heightened sensitivity to the semantic information embedded within the audio. These different features with distinct acoustic information from the same audio segment can be complementary. Hence, for the acoustic modality, our framework incorporates different HUBURT layers to extract low-level, mid-level, and high-level audio features, forming a unified representation of the original audio. For visual modality, we first crop and align faces of raw video in each frame using the OpenFace [2] toolkit. Then, we utilize various pre-trained models (such as MANet [21] and ResNet-50 [11]) to extract frame-level features and apply average encoding to compress them into video-level embeddings.

Then, these different acoustic and visual representations will be fused together with deep feature fusion frameworks based on attention-guided feature gathering (AFG) [20]. In addition, features of text modality, however, are proved to be underperforming comparing with acoustic and visual features in this task [19] so we do not take features from text encoders into consideration in our framework. The architecture of AFG module is shown in Figure 2, whose principle is as follows:

$$\boldsymbol{h}_i = \mathrm{Concat}(\boldsymbol{h}_i^1, \boldsymbol{h}_i^2, ..., \boldsymbol{h}_i^N) \tag{1}$$

$$\boldsymbol{\alpha}_i = \mathrm{Softmax}(\boldsymbol{h}_i^T \boldsymbol{W}_{\boldsymbol{\alpha}} + \boldsymbol{b}_{\boldsymbol{\alpha}}) \tag{2}$$

$$\hat{\boldsymbol{h}_i} = \boldsymbol{h}_i \boldsymbol{\alpha}_i \tag{3}$$

where $\boldsymbol{h}_i^1, \boldsymbol{h}_i^2, ..., \boldsymbol{h}_i^N \in \mathbb{R}^D$ are the aligned hidden states of sample i. $\boldsymbol{h}_i^T$ represents the transposed vector of $\boldsymbol{h}_i$, $\boldsymbol{\alpha}_i \in \mathbb{R}^N$ is the attention score that indices the importance of different features.$\boldsymbol{W}_{\boldsymbol{\alpha}} \in \mathbb{R}^{D \times N}$ and $\boldsymbol{b}_{\boldsymbol{\alpha}} \in \mathbb{R}^N$ are trainable attention matrix and bias. $\hat{\boldsymbol{h}_i} \in \mathbb{R}^D$ is the fused hidden state.

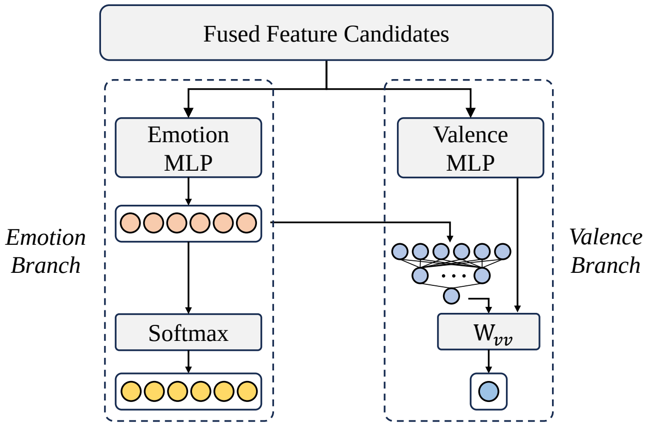

Figure 3: Joint decoding module of discrete and dimensional emotions. Fused feature candidates are the outputs of feature fusion module shown in Figure 1.

Three different feature fusion frameworks have been designed based on AFG [20], as depicted in Figure 1. In the first framework, we use AFG to fuse all features from acoustic and visual modalities in parallel. In the second framework, we fuse each acoustic representation with visual features to generate different audio-visual representations. These representations are then fused together, as they exhibit strong complementarity. While in the last framework, intra-modal fusion is firstly performed on the acoustic hidden states extracted by different layers of HUBURT [13] to create a unified acoustic representation, then inter-modal fusion is conducted based on the unified acoustic representation and visual representations.

2.2 Joint decoding of discrete and dimensional emotions

In our studies, we found there exists a relatively stable distribution between discrete and dimensional emotions [5, 30]. In other words, the discrete emotions can determine the dimensional emotions according to the following formula, where e_i represents discrete emotions while v_i represents dimensional emotions. k represents categories of emotions.

$$P(v_i) = \sum_{k=1}^{M} p(v_i|e_i = k)p(e_i = k) \tag{4}$$

Therefore, our research incorporates a branch for discrete emotions judgment of dimensional emotions into the network structure. This branch establishes a mapping from the discrete space to the dimensional space, as depicted in Figure 3. The prediction of discrete and dimensional emotions by our multi-task framework is as follows:

$$\hat{e_i} = \mathrm{Softmax}(\tilde{e_i}) = \mathrm{Softmax}(\hat{\boldsymbol{h}}_i \boldsymbol{W_e} + \boldsymbol{b_e}) \tag{5}$$

$$\hat{v_i} = \mathrm{Concat}(\tilde{v_i}, \tilde{v_i}^e)\boldsymbol{W_{vv}} + \boldsymbol{b_{vv}} \tag{6}$$

where $\hat{\boldsymbol{h}}_i \in \mathbb{R}^D$ is the fused feature, $\hat{\boldsymbol{e}}_i \in \mathbb{R}^C$ and $\hat{\boldsymbol{v}}_i \in \mathbb{R}^1$ are the estimated emotion and valence possibilities, respectively. $\boldsymbol{W_e} \in \mathbb{R}^{D\times C}$, $\boldsymbol{b_e} \in \mathbb{R}^{D\times C}$, $\boldsymbol{W_{vv}} \in \mathbb{R}^{2\times 1}$ and $\boldsymbol{b_{vv}} \in \mathbb{R}^{2\times 1}$ are trainable parameters. $\tilde{v_i} \in \mathbb{R}$ and $\tilde{v_i}^e \in \mathbb{R}$ are the estimated valence possibilities according to the fused state $\hat{\boldsymbol{h}}_i$ and emotion hidden state $\hat{\boldsymbol{e}}_i$ with trainable parameters $\boldsymbol{W_v} \in \mathbb{R}^{D\times 1}$, $\boldsymbol{b_v} \in \mathbb{R}^{D\times 1}$, $\boldsymbol{W_{ev}} \in \mathbb{R}^{C\times 1}$ and $\boldsymbol{b_{ev}} \in \mathbb{R}^{C\times 1}$, as follows:

$$\tilde{v_i} = \hat{\boldsymbol{h}}_i \boldsymbol{W_v} + \boldsymbol{b_v} \tag{7}$$

$$\tilde{v_i}^e = \mathrm{Tanh}(\tilde{\boldsymbol{e}}_i \boldsymbol{W_{ev}} + \boldsymbol{b_{ev}}) \tag{8}$$

During training, we use the cross-entropy (CE) loss as the classification loss, denoted as $\mathcal{L}_e$, for emotion prediction. The mean squared error (MSE) loss is adopted for valence prediction, denoted as $\mathcal{L}_v$. Moreover, we introduce dynamic weights to combine the two losses for better performance in the multi-task learning process. Inspired by the Uncertainty loss [6], we introduce uncertainty loss weighting to $\mathcal{L}_e$ and $\mathcal{L}_v$, whose principle is as follows:

$$\mathcal{L}_{ev} = \frac{1}{{\delta_1}^2}\mathcal{L}_e + \frac{1}{2{\delta_2}^2}\mathcal{L}_v + \log(1+\delta_1) + \log(1+\delta_2) \tag{9}$$

where δ_1 and δ_2 are trainable uncertainty weights. We improved the regular loss term to $\log(1+\delta_1)$ and $\log(1+\delta_2)$ to avoid effects caused by enormous negative weights.

3 RESULTS AND DISCUSSION

We have conducted several experiments to evaluate the effectiveness of the proposed multimodal framework.

3.1 Dataset and metric

In this research, we conduct experiments on MER 2023 dataset [19]. The dataset consists of 3373 labeled single-speaker video segments used as the training dataset. There are 411 and 412 unlabeled video segments for the test set in tracks 1 and 2, respectively. Same with the baseline [19], the combined metric of emotion classification and valence regression is chosen to evaluate the overall performance of discrete and dimensional emotions.

3.2 Implementations

We have employed several data augmentation techniques specifically designed for emotional data on the training dataset for MER-NOISE sub-challenge. For the audio modality, we add speaker-independent noise with 7 different signal-to-noise ratios (from 5dB to 11dB with a step of 1dB) from the *speech* subset of MUSAN [28] to simulate audio segments of various qualities. For the visual modality, we perform various image transformations to augment the data, including variations in brightness (e.g., solarize), changes in articulation (e.g., blur), alterations in position (e.g., rotate), and modifications in image content (e.g., cutout).

We also conduct decision-level fusion on predictions of multimodal emotion recognition systems with three fusion strategies. The fused predictions are as follows:

$$\hat{\boldsymbol{e}} = k_1\hat{\boldsymbol{e}_1} + k_2\hat{\boldsymbol{e}_2} + k_2\hat{\boldsymbol{e}_3} \tag{10}$$

$$\hat{\boldsymbol{v}} = k_1\hat{\boldsymbol{v}_1} + k_2\hat{\boldsymbol{v}_2} + k_3\hat{\boldsymbol{v}_3} \tag{11}$$

where $\hat{\boldsymbol{e}}_i$ and $\hat{\boldsymbol{v}}_i$ represents the emotion and valence prediction vectors from three different multimodal emotion recognition systems. $k_1, k_2, k_3 (k_1 + k_2 + k_3 = 1)$ represents the weighted factors of three systems respectively. The final decisions of emotion and valence are based on the posterior probability outputs of separate systems with different feature fusion encoders.

Table 1: Performance comparison of multimodal systems with different deep features. "Dis", "Dim", and "Com" denote the performance on discrete, dimensional, and combined metrics. JDEV is short for Joint Decoding for Emotion and Valence.

Feature Encoder		MR+RF		HL(18)+HL(19)+HL(20)		**HL(18)+HL(19)+HL(20)+MR+RF**							
Fusion Strategy		1		1		1		2		3		**1+2+3 fused**	
Decoder		Baseline	JDEV	Baseline	JDEV	Baseline	JDEV	baseline	JDEV	Baseline	JDEV	Baseline	**JDEV**
Train&Val	Dis(↑)	0.6085	0.6170	0.6995	0.7051	0.7743	0.7811	0.7769	0.7789	0.7735	0.7795	0.7865	**0.7936**
	Dim(↓)	1.2291	1.1890	0.9568	0.9297	0.6750	0.6176	0.6714	0.6339	0.6659	0.6315	0.6364	**0.6138**
	Com(↑)	0.3012	0.3198	0.4603	0.4727	0.6056	0.6267	0.6091	0.6204	0.6070	0.6216	0.6247	**0.6402**
MER-MULTI	Com(↑)	0.3072	0.3171	0.4997	0.5184	0.6675	0.6779	0.6550	0.6656	0.6618	0.6693	0.6787	**0.6846**
MER-NOISE	Com(↑)	0.2984	0.3074	0.4933	0.4934	0.6110	0.6178	0.6066	0.6166	0.6058	0.6125	0.6162	**0.6303**

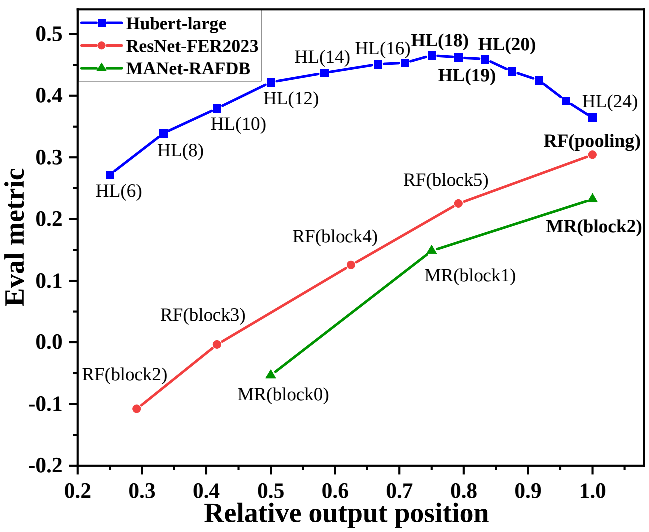

Figure 4: Performance comparison of unimodal systems using different deep features as input.

3.3 Performance comparison of unimodal systems

We compared the performance of different unimodal emotion recognition systems. Different unimodal systems utilizing different layers of the same pre-trained model as deep feature encoders have also been considered.

Figure 4 presents the performance diversity of different layers of audio and visual systems on Train&Val. HL(i) indicates using the output of i-th layer of HUBURT-large [13] model as acoustic feature. RF(j) and MR(j) indicate using the output of j-th block of ResNet-FER2013 [11] or MANet-RAFDB [21] as encoder output. Relative output position indicates the relative positions that the output layer in the whole pre-trained model (HUBURT-large [13], ResNet [11] or MANet [21]).

Among various acoustic features, HL(18), HL(19), HL(20) outperform others, confirming that the mid-level features from HUBERT-large [13] model is more suitable for emotion recognition. For visual modality, high-level features generated by the last block of ResNet [11] and MANet [21] outperform others. In general, acoustic features performs better than visual features in this task.

3.4 Performance comparison of multimodal systems

In this section, we perform multimodal fusion based on 3 well-performing acoustic features HL(18), HL(19), HL(20) and 2 visual features RF(pooling) and MR(block2). We conduct both intra-modal and inter-modal fusion. Three fusion strategies are conducted separately in inter-modal fusion. Then, the baseline[19] and the proposed joint decoding manner are utilized separately to obtain predictions for emotion and valence. The results are shown in Table 1.

The results indicate that features from different layers of the same model can be complementary, as the fusion metric improves by 1-2 percent points comparing with single-layer feature. Furthermore, the results demonstrate that incorporating features from different modalities significantly improves performance. The final multimodal system achieves a score of 0.6846 tested on MER-MULTI, which is a 16.6 percent improvement to the unimodal system.

Interestingly, we observed a significant gain on MSE loss for dimensional emotion regression when utilizing JDEV. This phenomenon suggests that the relatively reliable results of emotion classification can assist in improving the accuracy of dimensional valence regression by joint decoding.

Among the three proposed fusion strategies, the parallel fusion strategy 1 achieves the highest metric of 0.6779 on MER-MULTI. Additionally, when combining all three strategies on posterior probability level, we obtain an additional score gain of 0.67 percent points when tested on MER-MULTI sub-challenge.

4 CONCLUSIONS

In this paper, we propose a hierarchical audio-visual information fusion framework for recognizing both discrete and dimensional emotions. Three different feature fusion encoders are designed for deep feature fusion. In the decoding stage, we introduce a joint decoding structure for emotion classification and valence regression. In addition, a multi-task loss is also designed as optimizer for the whole process. Finally, by combining three different structures on posterior probability level, we obtain the final predictions of both emotion and valence. When tested on the dataset of MER 2023, our final system ranks third on the MER-MULTI sub-challenge.

5 ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grants No. 62171427.

REFERENCES

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477 [cs.CL]

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision*. 1–10. https://doi.org/10.1109/WACV.2016.7477553

[3] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal Multi-Task Learning for Dimensional and Continuous Emotion Recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (Mountain View, California, USA) *(AVEC '17)*. Association for Computing Machinery, New York, NY, USA, 19–26. https://doi.org/10.1145/3133944.3133949

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

[5] Huang-Cheng Chou, Chi-Chun Lee, and Carlos Busso. 2022. Exploiting Co-occurrence Frequency of Emotions in Perceptual Evaluations To Train A Speech Emotion Classifier. In *Proc. Interspeech 2022*. 161–165. https://doi.org/10.21437/Interspeech.2022-11041

[6] Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7482–7491. https://doi.org/10.1109/CVPR.2018.00781

[7] Jose Maria Garcia-Garcia, Victor M. R. Penichet, and Maria D. Lozano. 2017. Emotion Detection: A Technology Review. In *Proceedings of the XVIII International Conference on Human Computer Interaction*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3123818.3123852

[8] Hatice Gunes and Bjoern Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image & Vision Computing* 31, 2 (2013), 120–136.

[9] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 6–15. https://doi.org/10.1145/3462244.3479919

[10] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. arXiv:2109.00412 [cs.CL]

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Jonathan Herzig, Michal Shmueli-Scheuer, and David Konopnicki. 2017. Emotion Detection from Text via Ensemble Classification Using Word Embeddings. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 269–272. https://doi.org/10.1145/3121050.3121093

[13] 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

[14] Anthony Hu and Seth Flaxman. 2018. Multimodal Sentiment Analysis To Explore the Structure of Emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 350–358. https://doi.org/10.1145/3219819.3219853

[15] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7837–7851. https://aclanthology.org/2022.emnlp-main.534

[16] Anoop K, Deepak P, and Lajish V L. 2020. Emotion Cognizance Improves Health Fake News Identification. In *Proceedings of the 24th Symposium on International Database Engineering & Applications* (Seoul, Republic of Korea) *(IDEAS '20)*. Association for Computing Machinery, New York, NY, USA, Article 12, 10 pages. https://doi.org/10.1145/3410566.3410595

[17] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* 7 (2019), 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

[18] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational Transformer Network for Emotion Recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 29 (jan 2021), 985–1000. https://doi.org/10.1109/TASLP.2021.3049898

[19] Zheng Lian, Haiyang Sun, Licai Sun, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, et al. 2023. MER 2023: Multi-label Learning, Modality Robustness, and Semi-Supervised Learning. arXiv:2304.08981 [cs.CL]

[20] Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang. 2019. Conversational Emotion Analysis via Attention Mechanisms. In *Proc. Interspeech 2019*. 1936–1940. https://doi.org/10.21437/Interspeech.2019-1577

[21] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Mutual Affine Network for Spatially Variant Kernel Estimation in Blind Image Super-Resolution. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4076–4085. https://doi.org/10.1109/ICCV48922.2021.00406

[22] Huaishao Luo, Lei Ji, Yanyong Huang, Bin Wang, Shenggong Ji, and Tianrui Li. 2021. ScaleVLAD: Improving Multimodal Sentiment Analysis via Multi-Scale Fusion of Locally Descriptors. arXiv:2112.01368 [cs.CL]

[23] Mona Hafez Mahmoud. 2019. A Survey of Some Interdisciplinary Methods and Tools to Measure Learners' Emotions in Intelligent Tutoring Systems. In *2019 6th International Conference on Advanced Control Circuits and Systems (ACCS) and 2019 5th International Conference on New Paradigms in Electronics & information Technology (PEIT)*. 1–6. https://doi.org/10.1109/ACCS-PEIT48329.2019.9062885

[24] Mehdi Malekzadeh, Mumtaz Begum Mustafa, and Adel Lahsasna. 2015. A review of emotion regulation in intelligent tutoring systems. *Journal of Educational Technology & Society* 18, 4 (2015), 435–445.

[25] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. 2019. Audio-Visual Emotion Recognition in Video Clips. *IEEE Transactions on Affective Computing* 10, 1 (2019), 60–75. https://doi.org/10.1109/TAFFC.2017.2713783

[26] Keyur Patel, Dev Mehta, Chinmay Mistry, Rajesh Gupta, Sudeep Tanwar, Neeraj Kumar, and Mamoun Alazab. 2020. Facial Sentiment Analysis Using AI Techniques: State-of-the-Art, Taxonomies, and Challenges. *IEEE Access* 8 (2020), 90495–90519. https://doi.org/10.1109/ACCESS.2020.2993803

[27] Jouni Pohjalainen, Fabien Fabien Ringeval, Zixing Zhang, and Björn Schuller. 2016. Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands) *(MM '16)*. Association for Computing Machinery, New York, NY, USA, 670–674. https://doi.org/10.1145/2964284.2967306

[28] David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. arXiv:1510.08484

[29] Bogdan Vlasenko and Andreas Wendemuth. 2009. Processing affected speech within human machine interaction. In *10th Annual Conference of the International Speech Communication Association*. ISCA, Brighton, United Kingdom.

[30] Kexin Wang, Zheng Lian, Licai Sun, Bin Liu, Jianhua Tao, and Yin Fan. 2022. Emotional Reaction Analysis Based on Multi-Label Graph Convolutional Networks and Dynamic Facial Expression Recognition Transformer. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge* (Lisboa, Portugal) *(MuSe' 22)*. Association for Computing Machinery, New York, NY, USA, 75–80. https://doi.org/10.1145/3551876.3554810

[31] Shu-Lin Wang, I-En Chiang Honours, Alex Kuo, and Jing-Ya Lin. 2022. Mobile Emotion Healthcare System Applying Sentiment analysis. In *2022 IEEE International Conference on Big Data (Big Data)*. 2814–2820. https://doi.org/10.1109/BigData55660.2022.10021053

[32] 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. arXiv:2111.02735 [cs.CL]

[33] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. 2017. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Comput. Surv.* 50, 2, Article 25 (may 2017), 33 pages. https://doi.org/10.1145/3057270

[34] Jing Zhao and Wei-Qiang Zhang. 2022. Improving Automatic Speech Recognition Performance for Low-Resource Languages With Self-Supervised Models. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1227–1241. https://doi.org/10.1109/JSTSP.2022.3184480

[35] Zengqun Zhao and Qingshan Liu. 2021. Former-DFER: Dynamic Facial Expression Recognition Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*. Association for Computing Machinery, New York, NY, USA, 1553–1561. https://doi.org/10.1145/3474085.3475292

[36] Hengshun Zhou, Jun Du, Yuanyuan Zhang, Qing Wang, Qing-Feng Liu, and Chin-Hui Lee. 2021. Information Fusion in Attention Networks Using Adaptive and Multi-Level Factorized Bilinear Pooling for Audio-Visual Emotion Recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 29 (jul 2021), 2617–2629. https://doi.org/10.1109/TASLP.2021.3096037

[37] Hengshun Zhou, Debin Meng, Yuanyuan Zhang, Xiaojiang Peng, Jun Du, Kai Wang, and Yu Qiao. 2019. Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 562–566. https://doi.org/10.1145/3340555.3355713