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

Multi-Modal Multi-Task Affective States Recognition Based on Label Encoder Fusion

Maxim Markitantov[1,*], *Elena Ryumina*[1,*], *Heysem Kaya*[2], *Alexey Karpov*[1]

[1]St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Russia
[2]Department of Information and Computing Sciences, Utrecht University, The Netherlands

m.markitantov@yandex.ru, ryumina_ev@mail.ru, h.kaya@uu.nl, karpov@iias.spb.su

Abstract

Despite recent advances in multi-modal approaches, recognizing the full range of human affective states, including emotions and sentiments, remains challenging due to complex interactions between different modalities and the hierarchical nature of affective states. This work presents a novel approach for multi-modal multi-task emotion and sentiment recognition that integrates audio, video, and text data. We introduce a Label Encoder Fusion Strategy, which produces and processes uni-modal emotion and sentiment predictions, which are used alongside modality-specific features during the fusion process to provide additional contextual information. We conduct elaborate multi-corpus experiments on the RAMAS, MELD, and CMU-MOSEI corpora. The proposed approach achieves state-of-the-art performance in both affective tasks. On MELD, we achieve a macro F1 (MF) of 40.9% and 67.02% for emotion and sentiment recognition. On CMU-MOSEI, the mean MF is 62.30% and MF is 62.00% for the same tasks.

Index Terms: emotion recognition, sentiment recognition, multi-modal fusion, label encoder, multimodal paralinguistics

1. Introduction

Affective computing has gained significant attention in recent years, driven by the need for more natural human-computer interaction, with applications ranging from mental health monitoring [1] to customer feedback analysis [2]. Emotion and sentiment recognition from multi-modal data is a crucial aspect of this research domain [3, 4], as human affects are inherently expressed through multiple channels, including the audio, video and text modalities [5, 6, 7]. Traditional uni-modal methods often fail to capture the full range of affective expressions, resulting in non-optimal performance in real-world applications [6, 8]. Therefore, multi-modal affective states recognition has become a promising direction that takes advantage of the complementary nature of different data sources [6, 7].

Recent advances in deep learning, particularly transformer-based [9, 10, 11] and graph-based [12, 13] architectures, have significantly improved affective states recognition performance. Chauhan et al. [14] and Sangwan et al. [15] employed autoencoders and cross-modal attention methods to capture inter-modal dependencies. Based on this, Zhang et al. [10] introduced intra- and inter-modal attention, while Hwang et al. [16] used pseudo-labels for affective states recognition. Tu et al. [13] proposed a method involving speaker-addressee interactions and contrastive learning for emotion shift detection. Van et al. [12] developed a method for integrating heterogeneous graphs and hypergraphs for contextual representation. Finally, Zheng et al. [11] introduced a method for optimizing intra- and inter-task dynamics using mutual information maximization. For each modality, several approaches have been proposed, enabling more effective feature extraction from speech [17, 18, 19] facial expressions [20, 21, 22], and textual content [23, 24]. However, existing multi-modal approaches often suffer from inefficient modality fusion and insufficient utilization of inter-modal relationships. Moreover, although most approaches focus on single-task classification, the real-world applications require multi-task solutions that can jointly perform emotion and sentiment recognition in various modalities and languages [25].

In Machine Learning (ML) literature, the use of predictions or probabilities from an ML model as an input to another model is known as *stacking* [26]. This approach is successfully used in multi-modal affective computing to improve prediction performance [27], especially with Random Forests, which provide out-of-bag predictions as a robust estimate of the test set performance. Further in this direction, Han et al. [28] proposed combining model's predictions with its input feature vector for a subsequent model, in an approach called *strength modeling* for uni- and multi-modal affect prediction.

This work introduces a novel approach for affective states recognition, including emotions and sentiments, which integrates state-of-the-art (SOTA) models for processing audio, video, and text modalities, and a Label Encoder Fusion Strategy (LEFS). The latter is based on transformer blocks and models cross-modal interactions, ensuring a comprehensive integration of all modalities, while additional averaging balances uni- and multi-modal predictions, enhancing model's robustness.

The growing need for universal and integrated systems emphasizes the relevance of this work. The proposed approach provides increased flexibility and broader applicability by supporting different types of content and multiple data sources that distinguish it from existing solutions. By combining advanced models with an innovative fusion, this approach advances multi-modal multi-task affective states recognition and provides a flexible framework for multi-modal learning and multi-lingual affective states recognition.

2. Proposed approach

The pipeline of the proposed approach is presented in Figure 1. It includes data pre-processing, feature extraction using audio-based, image-based, and text-based models as well as multi-modal fusion, which is performed using the proposed Label Encoder Fusion Strategy with Averaging (LEFSA).

2.1. Audio-based models

At the initial step, we segment the data into short overlapping 3-second segments with a 2-second step size that helps capture lo-

*These authors contributed equally.

10.21437/Interspeech.2025-2060

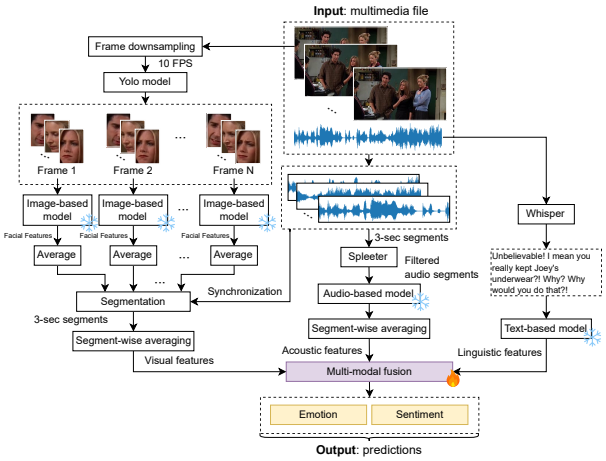

Figure 1: *General pipeline of the approach proposed.*

cal emotional variations. After that, we pre-process the speech signal by filtering the audio signal from background noise using Spleeter from deezer[1]. To extract acoustic features, we use two pre-trained SOTA models for Speech Emotion Recognition (SER). The first model is a pruned Wav2Vec2 model[2] trained on the MSP-Podcast corpus [29]. This model is trained for dimensional SER and predicts arousal, dominance, and valence, representing different aspects of emotional expression in audio.

The ExHuBERT model[3] is an enhanced version of HuBERT with Backbone Block Expansion with extra zero-initialized linear layer and skip connections [19]. This model was trained on the EmoSet++ corpus, a comprehensive multilingual, multi-cultural speech emotion corpus that is a combination of 37 corpora of acted and spontaneous emotional speech, which enhances the model's ability to capture controlled and natural emotional expressions. This model predicts two arousal levels (high, low) and three valence levels (positive, neutral, negative), aggregated into six classes (Disgust, Neutral, Kind, Anger, Surprise, and Joy). We extract deep acoustic features from the pooled states of the last transformer layers with the size of 149×1024 of both models.

2.2. Image-based models

As preprocessing, we use the YOLO model[4] for face detection on thinned frames (10 FPS). Since research corpora contain multiple faces in each frame, the final feature vector is computed as the average of all encoded face representations. To extract visual features from frames, we use two SOTA models trained on the large-scale AffectNet corpus, which is designed to recognize six basic emotions and a neutral state. The EmoAffectNet [20][5] model is based on the ResNet-50 architecture. This model was trained using multiple augmentation methods, including Mixup [30], labels smoothing [31], random affine transformations, and contrast adjustment.

The ResEmoteNet [22] model[6] is based on the ResNet architecture, but also includes additional convolutional blocks and a squeeze and excitation block. Due to the modified ResNet structure, it outperforms all SOTA models developed on the AffectNet corpus. Thus, we extract deep visual features for each frame with a size of 512 and 256 using EmoAffectNet and ResEmoteNet models, respectively. After that, we segment the sequence of visual-representations into short overlapping 3-second segments with a 2-second step size according to audio data (10 FPS×3 = 30 frames).

2.3. Text-based models

For the Multimodal EmotionLines Dataset (MELD) [32] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [33] corpora, we use originally provided transcriptions in English. In the case of the Russian Acted Multimodal Affective Set (RAMAS) corpus [34], we automatically extracted transcriptions from Russian speech using Whisper [35] [7]. Unlike the emotional models used for extracting audio and video features, we use multi-lingual rather than emotional models to extract linguistic features from textual data. This choice is due to the fact that emotional text models are usually trained for a single language, while this work is based on speech transcriptions in both English and Russian. To extract textual (or linguistic) features, we use two SOTA models. The XLM-RoBERTa [23] model[8] is the transformer-based model designed to learn internal representations for 100 languages. The vocabulary size of this model is above 250K words.

The jina-embeddings-v3 [24] (or JINA) model[9] is based on XLM-RoBERTa, and can handle long input sequences due to Rotary Position Embeddings. The model specializes in internal representations for 30 languages, and the vocabulary size is the same as the XLM-RoBERTa model. Across all the corpora, samples contain an average of 48 tokens. Thus, we extract deep linguistic features of 48×768 and 48×1024 using the XLM-RoBERTa and JINA models, respectively.

2.4. Multi-modal multi-task affective states recognition

In this paper, we propose three different fusion strategies:

- **Baseline Fusion Strategy (BFS)**. This strategy involves an interaction of three feature representations (acoustic, visual, and linguistic) equally through the cross-modal attention mechanism. In the initial stage, the uni-modal features are downsampled to the same size of 30 × 256 using a set of convolutional and normalization layers. These features are processed using six transformer blocks that perform a cross-modal attention. The resulting feature representations contain information about the interaction of each modality with the other two modalities. The representations based on the same query are combined with each other, from which average values are calculated. They are aggregated later with their processed versions by fully-connected layers. The resulting multi-modal features are passed through two fully-connected layers to obtain emotion and sentiment predictions. This is the basic strategy used later to compare with other strategies. This model has 6.6M parameters.
- **LEFS**. This strategy is based on the previous strategy. It involves a Label Encoder, which produces emotion and sentiment predictions of the corresponding modality and encodes them into new feature representations with the size of 30 × 256. A predictions upsampling is performed using the

[1]https://github.com/deezer/spleeter
[2]https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
[3]https://huggingface.co/amiriparian/ExHuBERT
[4]https://github.com/lindevs/yolov8-face
[5]https://huggingface.co/ElenaRyumina/face_emotion_recognition
[6]https://github.com/ArnabKumarRoy02/ResEmoteNet
[7]https://huggingface.co/openai/whisper-base
[8]https://huggingface.co/FacebookAI/xlm-roberta-base
[9]https://huggingface.co/jinaai/jina-embeddings-v3

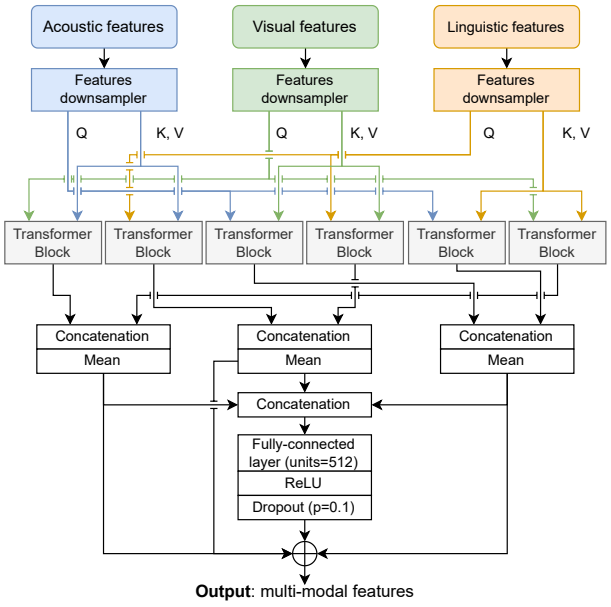

Figure 2: *Label Encoder Fusion Strategy with Averaging.*

deconvolution operation. The obtained features are summarized with the query and the value of each modality before the cross-modal attention:

$$\hat{Q}_X = Q_X + LabelEncoder(Q_X), \quad (1)$$

$$\hat{K}_X = K_X + LabelEncoder(K_X), \quad (2)$$

where Q_X and K_X are the query and key of the X modality. This model has 7.1M parameters.

- **LEFSA**. This strategy is based on the previous strategy. However, in addition to deep features, this strategy utilizes emotion and sentiment predictions of each modality obtained from the Label Encoder (see Figure 2). These uni-modal predictions are combined with the multi-modal predictions and then averaged. This model has 7.1M parameters.

3. Experiments

3.1. Research corpora

We use three multi-modal corpora for affective states modeling: RAMAS [34], CMU-MOSEI [33], and MELD [32]. RAMAS consists of audio-visual Russian speech and motion-capture data from dyadic interactions on various topics. MELD comprises dialogues of the Friends TV-series that were labeled by three annotators. CMU-MOSEI includes YouTube monologues on 250 topics in English. All corpora include audio, video, and text modalities; the labels are mapped into six emotion classes (Anger, Disgust, Happiness, Sadness, Fear, and Surprise) with Neutral and three sentiment classes (Negative, Neutral, and Positive). In order to perform a multi-corpus research and combine CMU-MOSEI with the other two corpora, we scale multi-label emotion labels of CMU-MOSEI as follows:

$$\hat{t}_i = \frac{t_i}{\sum_{i=1}^{C} t_i}, \quad (3)$$

where i varies from 1 to C, which is the total number of emotion classes, t_i is the value of the target emotion class.

The Train, Devel and Test subsets are pre-defined in all corpora. Emotion-sentiment distributions are shown in Figure 3.

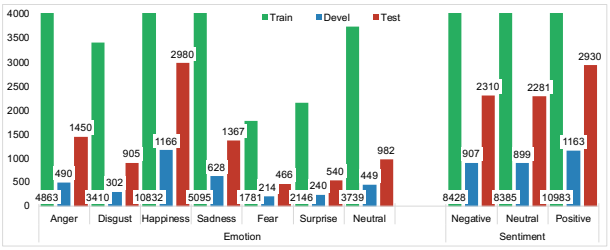

Figure 3: *Emotion and sentiment distributions in the combined RAMAS, MELD, and CMU-MOSEI corpora.*

3.2. Experimental setup

Most SOTA approaches focus on single-corpus research, leading to model overfitting on data from a single domain, while their performance on new data is degraded. In this paper, we focus on the development of multi-corpus models, which enable creation of models that are capable to cover broader domains (dialogues and monologues, scripted and spontaneous speech, English and Russian languages) and improve their generalization ability to new data.

To evaluate the proposed approach, we use macro F1-score (MF) for both emotion and sentiment recognition. However, in the case of CMU-MOSEI, we compute the performance measure as in multi-label emotion recognition; we use a macro F1-score averaged between binary positive and negative emotion classes (mMF). Macro measures reflect the actual performance of the model for both minority and majority classes. However, for comparison with SOTA, we use weighted measures such as accuracy (A) and weighted F1-score (WF) for RAMAS and MELD), and weighted accuracy (WA) and mean weighted F1-score (mWF) for CMU-MOSEI, despite their bias toward majority classes. We use the sum of two weighted Categorical Cross-Entropy (CCE) loss functions for emotion and sentiment recognition.

We experimented with various number of transformer blocks, fully-connected layers and their interaction, as well as the label encoder. We also experimented with various combinations of modalities and features. All the models were trained using the AdamW optimizer with the learning rate of $1e-5$, $\beta = 0.9$, $\beta = 0.999$, and $\epsilon = 1e-8$ for 75 epochs.

3.3. Experimental results and discussion

The experiments were conducted in two stages (see Table 1). Firstly, we determined the optimal combination of uni-modal features with BFS (models 1-8). Then we experimented with the best combination of features and two additional fusion strategies (LEFS and LEFSA) by models 9 and 10. Depending on the combination of features used, the average performance measure varied from 53.18 to 59.18% for all the corpora. The visual features extracted using EmoAffectNet have shown better performance than ResEmoteNet features. EmoAffectNet is probably trained with multiple augmentation methods that enhance its generalizability. The linguistic features extracted using JINA outperformed those of RoBERTa. This is probably because JINA covers fewer languages (30 vs. 100), which may increase its effectiveness in particular languages. Both types of acoustic features show similar results. However, Wav2Vec2

Table 1: *Experimental results obtained for the multi-modal affective states recognition. Performance measure superscript shows the number of classes. The average value is an integral score calculated from the sum of the performance measures in a row.*

					RAMAS		MELD		CMU-MOSEI		
					Emotion	Sentiment	Emotion	Sentiment	Emotion	Sentiment	
ID	**Audio**	**Video**	**Text**	**Fusion**	MF^7	MF^3	MF^7	MF^3	mMF^6	MF^3	**Average**
1	Wav2Vec2	EmoAffectNet	JINA	BFS	60.57	65.02	38.56	65.94	62.46	**62.56**	59.18
2	ExHuBERT	EmoAffectNet	JINA	BFS	62.35	64.02	38.40	63.93	62.10	62.32	58.85
3	Wav2Vec2	EmoAffectNet	RoBERTa	BFS	57.64	**67.14**	35.48	65.03	61.58	60.19	57.84
4	ExHuBERT	EmoAffectNet	RoBERTa	BFS	60.54	64.70	35.05	63.78	60.78	59.25	57.35
5	Wav2Vec2	ResEmoteNet	JINA	BFS	55.88	59.93	37.75	63.81	62.22	63.69	57.21
6	ExHuBERT	ResEmoteNet	JINA	BFS	57.21	61.40	38.25	64.06	61.11	60.81	57.14
7	ExHuBERT	ResEmoteNet	RoBERTa	BFS	49.62	54.44	32.56	62.15	59.88	60.47	53.19
8	Wav2Vec2	ResEmoteNet	RoBERTa	BFS	52.29	54.87	34.08	61.65	58.79	57.42	53.18
9	Wav2Vec2	EmoAffectNet	JINA	LEFS	61.38	66.57	39.79	65.70	**62.79**	61.59	59.64
10	Wav2Vec2	EmoAffectNet	JINA	LEFSA	**62.52**	64.96	**40.09**	**67.02**	62.30	62.00	**59.81**

performs better in a combination with JINA and EmoAffectNet, but not ExHuBERT, which is probably due to the larger training data (250 h vs. 150 h) and its domain (conversational speech). Therefore, the first stage of the experiments showed that the combination of features extracted using Wav2Vec2, EmoAffectNet, and JINA outperforms the other combinations in multi-modal, multi-task, and multi-corpus setups.

Table 2: *Comparison of the proposed approach with SOTA multi-modal approaches. Superscripts at the performance measures show the number of classes. ST and MT refer to the single-task and multi-task affective states recognition.*

Approach	**Corpus**	**Setup**	**Emotion**		**Sentiment**	
			A^7	WF^7	A^3	WF^3
Ours	RAMAS	MT	**68.99**	**67.79**	**84.11**	**84.02**
Zhang et al. [10]	MELD	MT	41.17	41.22	67.33	67.21
Van et al. [12]		ST	66.28	65.69	–	–
Hwang et al. [16]		ST	66.70	65.93	–	–
Tu et al. [13]		ST	**67.85**	**67.02**	–	–
Ours		MT	62.30	59.79	**69.20**	**69.02**
			mWA^6	mWF^6	A^2	WF^2
Chauhan et al. [14]	CMU-MOSEI	MT	62.97	79.02	80.37	78.23
Sangwan et al. [15]		MT	63.16	79.06	80.15	78.30
Hwang et al. [16]		ST	–	–	**87.40**	**87.30**
Zheng et al. [11]		ST	–	–	85.90	86.00
Ours		MT	**64.78**	**79.06**	84.83	84.90

When BFS is replaced by alternative strategies (LEFS and LEFSA), a performance improvement is also observed. In particular, LEFS shows a better performance for the RAMAS and MELD corpora, while a performance degradation is observed for CMU-MOSEI, which may indicate the multi-label annotation nature of the latter corpus. In contrast, the use of LEFSA leads to performance improvements for the RAMAS and MELD corpora compared to BFS, while the performance measures for the CMU-MOSEI corpus remain unchanged. Therefore, the use of LEFSA improves the intra- and inter-modal relationships between the modalities and tasks, leading to better outcomes.

We compare the proposed approach with SOTA multi-task and single-corpus approaches (see Table 2). It is important to note that our models are trained in a multi-corpus and multi-lingual setup that distinguishes our approach from the rest. In general, the multi-task approaches show lower performance compared to the single-task approaches. We provide the baseline multi-modal results for RAMAS. For MELD, our approach outperforms the multi-task approach [10] by A=21.13% and WF=18.57% in emotion recognition and by A=1.87% and WF=1.81% in sentiment recognition. However, our approach is inferior to the single-task approach [13] by A=5.55% and WF=7.23% in emotion recognition. A similar trend is observed for CMU-MOSEI. In the sentiment recognition task, our approach outperforms the multi-task approach [15] by A = 4.68% and WF = 6.60%, while being inferior to the single-task approach [16] by A=2.57% and WF=2.40%. Therefore, the results demonstrate the superiority of the approach proposed over SOTA approaches, offering an improved intra- and inter-modal interaction, as well as multi-task and multi-lingual capabilities that enhance the model's generalizability to new data.

In addition, we report the computational cost of our approach. The real-time factor for processing a 1-second multi-modal data using Wav2Vec2, Yolo, EmoAffectNet, Whisper, and JINA models is 0.38 sec using GPU (GeForce RTX 3080) and 0.52 sec using CPU (Intel i9).

4. Conclusions

In this paper, we proposed the multi-modal multi-task approach for affective states recognition (emotion and sentiment) that effectively integrates audio, video and text data by the Label Encoder Fusion Strategy. This strategy produces and processes uni-modal emotion and sentiment predictions, providing additional contextual information, which are used alongside the modality-specific features during the information fusion process. The experimental results have highlighted the advantages of the specific combination of features extracted by Wav2Vec2, EmoAffectNet, and JINA that provide the best performance. Although the proposed approach has outperformed the SOTA multi-task approaches, it remained slightly inferior to some of the SOTA single-task approaches. However, these performance disadvantages are compensated for by high recognition rates and shared feature representations that improve generalization and improve the use of data in related tasks. These findings can highlight the potential of multi-modal and multi-task learning to advanced emotion and sentiment recognition by utilizing richer cross-modal relationships and a better adaptation to diverse corpora.

5. Acknowledgements

This research was supported by the Russian state research № FFZF-2025-0003.

6. References

[1] M. Niu, A. Romana *et al.*, "Capturing mismatch between textual and acoustic emotion expressions for mood identification in bipolar disorder," in *Proc. of the Interspeech*, Dublin, Ireland, Aug 2023, pp. 1718–1722.

[2] E. Wemmer, S. Labat, and R. Klinger, "Emoprogress: Cumulated emotion progression analysis in dreams and customer service dialogues," in *Proc. of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, Torino, Italia, May 2024, pp. 5660–5677.

[3] J. Li, S. Wang *et al.*, "Context-aware multimodal fusion for emotion recognition," in *Proc. of the Interspeech*, Incheon, Korea, Sep 2022, pp. 2013–2017.

[4] Y. Zhi, J. Li *et al.*, "A fine-grained tri-modal interaction model for multimodal sentiment analysis," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Apr 2024, pp. 5715–5719.

[5] R. W. Picard, *Affective computing*. MIT press, 2000.

[6] S. Poria, E. Cambria *et al.*, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[7] Y. Wang, W. Song *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[8] M. Shamsi, L. Gauder, and M. Tahon, "The conilium proposition for odyssey emotion challenge leveraging major class with complex annotations," in *The Speaker and Language Recognition Workshop (Odyssey)*, Quebec City, Canada, Jun 2024.

[9] M. A. Pastor, D. Ribas *et al.*, "Cross-corpus speech emotion recognition with hubert self-supervised representation," in *Proc. of the IberSPEECH 2022*, Granada, Spain, Nov 2022, pp. 76–80.

[10] Y. Zhang, J. Wang *et al.*, "A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations," *Information Fusion*, vol. 93, pp. 282–301, 2023.

[11] Y. Zheng, J. Gong *et al.*, "Djmf: A discriminative joint multi-task framework for multimodal sentiment analysis based on intra-and inter-task dynamics," *Expert Systems with Applications*, vol. 242, p. 122728, 2024.

[12] C. T. Van, T. V. T. Tran *et al.*, "Effective context modeling framework for emotion recognition in conversations," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr 2025, p. in print.

[13] G. Tu, F. Xiong *et al.*, "A persona-infused cross-task graph network for multimodal emotion recognition with emotion shift detection in conversations," in *Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Washington DC, USA, Jul 2024, pp. 2266–2270.

[14] D. S. Chauhan, M. S. Akhtar *et al.*, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov 2019, pp. 5647–5657.

[15] S. Sangwan, D. S. Chauhan *et al.*, "Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis," in *Proc. of the International Conference on Neural Information Processing (ICONIP)*, Sydney, NSW, Australia, Dec 2019, pp. 662–669.

[16] Y. Hwang and J.-H. Kim, "Easum: Enhancing affective state understanding through joint sentiment and emotion modeling for multimodal tasks," in *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, Hawaii, Jan 2024, pp. 5668–5678.

[17] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun 2023, pp. 1–5.

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

[19] S. Amiriparian, F. Packań *et al.*, "Exhubert: Enhancing hubert through block extension and fine-tuning on 37 emotion datasets," in *Proc. of the Interspeech*, Kos, Greece, Sep 2024, pp. 2635–2639.

[20] E. Ryumina, D. Dresvyanskiy, and A. Karpov, "In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study," *Neurocomputing*, pp. 435–450, 2022.

[21] W. Zhang, F. Qiu *et al.*, "Transformer-based multimodal information fusion for facial expression analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Seattle, WA, Jun 2022, pp. 2428–2437.

[22] A. K. Roy, H. K. Kathania *et al.*, "Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition," *IEEE Signal Processing Letters*, pp. 1–5, 2024.

[23] A. Conneau, K. Khandelwal *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul-Aug 2019, p. 8440–8451.

[24] S. Sturua, I. Mohr *et al.*, "jina-embeddings-v3: Multilingual embeddings with task lora," *arXiv preprint arXiv:2409.10173*, pp. 1–20, 2024.

[25] P. Bujnowski, B. Kuzma *et al.*, "Samsemo: New dataset for multilingual and multimodal emotion recognition," in *Proc. of the Interspeech*, Kos, Greece, Sep 2024, pp. 2925–2929.

[26] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[27] H. Kaya, F. Gurpinar, and A. Ali Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Honolulu, Hawaii, Jul 2017.

[28] J. Han, Z. Zhang *et al.*, "Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals," *Image and Vision Computing*, vol. 65, pp. 76–86, 2017.

[29] J. Wagner, A. Triantafyllopoulos *et al.*, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[30] H. Zhang, M. Cissé *et al.*, "Mixup: Beyond empirical risk minimization," in *Proc. of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 2018, pp. 1–13.

[31] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–10, 2019.

[32] S. Poria, D. Hazarika *et al.*, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul 2019, pp. 527–536.

[33] A. B. Zadeh, P. P. Liang *et al.*, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, Jul 2018, pp. 2236–2246.

[34] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "Ramas: Russian multimodal corpus of dyadic interaction for affective computing," in *Proc. of the International Conference on Speech and Computer (SPECOM)*, Leipzig, Germany, Sep 2018, pp. 501–510.

[35] A. Radford, J. W. Kim *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. of the International Conference on Machine Learning*, Kalākaua Avenue Honolulu, HI, Jul 2023, pp. 28 492–28 518.