# Multi-Microphone and Multi-Modal Emotion Recognition in Reverberant Environment

Ohad Cohen
*Faculty of Engineering*
*Bar-Ilan University*
Ramat-Gan, Israel
ohad.cohen@biu.ac.il
0009-0002-4707-881X

Gershon Hazan
*Faculty of Engineering*
*Bar-Ilan University*
Ramat-Gan, Israel
gershon.hazan@biu.ac.il

Sharon Gannot
*Faculty of Engineering*
*Bar-Ilan University*
Ramat-Gan, Israel
sharon.gannot@biu.ac.il
0000-0002-2885-170X

***Abstract*—This paper presents a Multi-modal Emotion Recognition (MER) system designed to enhance emotion recognition accuracy in challenging acoustic conditions. Our approach combines a modified and extended Hierarchical Token-semantic Audio Transformer (HTS-AT) for multi-channel audio processing with an $R(2+1)$D Convolutional Neural Networks (CNN) model for video analysis. We trained and evaluated our proposed method on a reverberated version of the Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset using synthetic and real-world Room Impulse Responses (RIRs). Our results demonstrate that integrating audio and video modalities yields superior performance compared to uni-modal approaches, especially in challenging acoustic conditions. Moreover, we show that the multimodal (audiovisual) approach that utilizes multiple microphones outperforms its single-microphone counterpart.**

***Index Terms*—Emotion recognition, multi-modal, reverberant conditions, audio transformer**

## I. Introduction

Emotion Recognition (ER) is a crucial component in human-computer interaction, with applications ranging from healthcare to customer service. Humans naturally express emotions across multiple modalities, including facial expressions, language, speech, and gestures. Accurately modeling the interactions between these modalities, which contain complementary and potentially redundant information, is essential for effective emotion recognition. Most existing studies primarily focus on uni-modal emotion recognition, concentrating on either text, speech, or video [1], [2], [3]. Although significant advancements in single-modal emotion recognition have been demonstrated, these models often fall short in complex scenarios since they do not utilize the inherently multi-modal nature of emotional expression. Moreover, research on jointly employing multi-modal and multi-microphones for ER is relatively scarce. Previous works have made significant strides in MER. Studies such as [4], [5], [6], [7] have developed systems that simultaneously analyze visual and acoustic data. In [8], researchers presented an unsupervised MER feature learning approach incorporating audio-visual and textual information. These studies often overlooked the challenges posed by real-world acoustic conditions, particularly reverberation and noise, which can significantly impact the performance of audio-based emotion recognition. Feature selection is vital in designing effective MER systems. For acoustic features, log-mel filterbank energies and log-mel spectrograms have been widely adopted [9], [10]. In the video domain, various deep learning architectures such as VGG16 [11], I3D [12], and FaceNet [13] have been employed, along with facial features like landmarks and action units extracted using tools like OpenFace [14]. For the text modality, Global Vectors for Word Representation (GloVe) [15] have been frequently used [14], [16], [17].

Despite these advancements, a gap remains in addressing the challenges posed by reverberant and noisy environments. Real-world acoustic conditions can significantly alter speech signals, potentially degrading the performance of audio-based emotion recognition systems. Moreover, the integration of multi-channel audio processing techniques with video analysis for emotion recognition in such challenging conditions has not been thoroughly explored.

This work addresses these limitations by proposing a MER with multi-channel audio that outperforms solutions solely based on single-channel audio. We propose a novel approach that combines two state-of-the-art architectures for audio-visual emotion recognition. The multi-channel extension of the HTS-AT architecture for the audio modality [18] and the $R(2+1)$D CNN model [19] for the video modality. We use a reverberated version of the RAVDESS [20] dataset to analyze the proposed scheme's performance. Reverberation was added by convolving the speech utterances with real-life RIRs drawn from the Acoustic Characterisation of Environments (ACE) challenge dataset [21]. The code of the proposed method is available.[1]

## II. Problem Formulation

Denote the two modalities as $M = \{\text{video}, \text{audio}\}$ and the set of emotions as:

$$E = \{\text{happy}, \text{calm}, \text{sad}, \text{angry}, \ldots \\ \text{neutral}, \text{fearful}, \text{disgust}, \text{surprised}\}. \quad (1)$$

Let $v(t)$ be the video signal and $s(t)$ the anechoic audio signal, with $t$ the time index. An array of $C$ microphones captures

The project has received funding from the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreement No. 871245; and from the Audition Project, Data Science Program, Council of Higher Education, Israel.

[1] https://github.com/OhadCohen97/Multi-Microphone-Multi-Modal-Emotion-Recognition-in-Reverberant-Environments.

the audio signal after propagating in the acoustic environment. The signals, as captured by the microphones, are given by:

$$y_i(t) = \{s * h_i\}(t),\ i = 1, 2, \ldots, C, \tag{2}$$

where $h_i(t),\ i = 1, 2, \ldots, C$, are the RIRs from the source to the $i$th microphone. The feature embeddings for each modality are denoted $f_v$ and $f_s$, respectively. This study aims to classify the utterance to one of the emotions using the available information and utilizing the relations between the feature embeddings of both modalities:

$$M\left\{v(t), \{y_i(t)\}_{i=1}^{C}\right\} \Rightarrow f_v \oplus f_s \Rightarrow E, \tag{3}$$

where $\oplus$ stands for late fusion concatenation.

## III. Proposed Model

Our proposed MER architecture combines two powerful models: the modified and extended HTS-AT [18] for multi-channel audio processing and the $R(2+1)$D model [19] for video analysis. These uni-modal models are integrated to create a robust multi-modal system for emotion recognition in challenging acoustic conditions.

The input features of the models are the mel-spectrograms for the audio track and the raw RGB facial images for the visual track. For the audio modality, we followed the same preprocessing procedure as in [18] and used the SpecAugment Library [22] to augment the mel-spectrograms. For augmenting the video modality, we used the TorchVision Library [23].
**Audio:** The extended multi-channel HTS-AT model addresses the integration of multi-microphone information, employing the Swin-Transformer architecture [24],[2] a variant of the Vision Transformer (ViT) [25] architecture. The architecture is also applicable to the single-microphone configurations, namely $C = 1$. The model's architecture consists of four groups, each comprising Swin-Transformer blocks with varying depths. In addition, the model uses a hierarchical structure and windowed attention mechanism to efficiently process mel-spectrograms, which serve as our audio feature extractor. We use the two multi-channel variants with the modified HTS-AT module, as proposed in [18]: 1) Patch-Embed Summation - the mel-spectrogram of each channel is processed through a shared Patch-Embed layer, after which the outputs are summed across channels; and 2) Average mel-spectrograms - mel-spectrograms from multiple channels are averaged before being fed into the model. More details can be found in [18].
**Video:** For video feature extraction, we employ the pre-trained $R(2+1)$D model, an 18-layer ResNet-based architecture designed for action recognition. The $R(2+1)$D model decomposes the 3D convolutions into separate spatial (2D) and temporal (1D) convolutions, which allows it to effectively capture both spatial and temporal features in the video data. Fig. 1 presents the integration of the two modalities.
**Feature Concatenation:** The feature embeddings are extracted from the extended multi-channel HTS-AT and the $R(2+1)$D models, followed by concatenation to create a unified multi-modal representation. This combined feature vector captures audio and visual cues relevant to emotion recognition. The concatenated features are then passed through two fully connected layers for final classification. These layers learn to interpret the combined audio-visual features and to map them to emotion categories. The output of the final layer corresponds to the predicted emotion class. This integrated scheme ensures that the multi-channel audio and visual data are effectively processed and leveraged. This allows the model to capture and utilize complementary information from both modalities, thus achieving improved ER accuracy.

## IV. Experimental Study

This section outlines the experimental setup and describes the comparative analysis between the proposed scheme and a baseline method.

### A. Datasets

Our work utilized the RAVDESS dataset for emotion recognition. This dataset includes 24 actors, equally divided between male and female speakers, each delivering 60 English sentences. Hence, there are 1440 audio-video pairs representing eight different emotions ('sad,' 'happy,' 'angry,' 'calm,' 'fearful,' 'surprised,' 'neutral,' and 'disgust'). All utterances are pre-transcribed. Therefore, the emotions are expressed more artificially compared to spontaneous conversation. The RAVDESS dataset is balanced across most classes except for the neutral class, which has a relatively small number of utterances. We used an *actor-split* approach, dividing the data into 80% training, 10% validation, and 10% test sets, ensuring no actor appears in more than one split. As a result, model accuracy may be lower than reported in some prior works because the test set includes actors not seen during fine-tuning. As publicly available multi-microphone datasets for Speech Emotion Recognition (SER) do not exist, we generated our own dataset. We used synthesized RIRs to fine-tune the multi-channel experiment model. We employed the 'gpuRIR' Python package[3] to simulate reverberant multi-channel microphone signals (setting the number of microphones to $C = 3$). Each clean audio sample from the RAVDESS dataset was convolved with distinct multi-channel RIRs, resulting in 1440 3-microphone audio samples. The associated video data is unaffected by reverberation. We simulated rooms with lengths and widths uniformly distributed between 3 m and 8 m, maintaining a constant height of 2.9 m and an aspect ratio between 1 and 1.6. We randomly positioned the sound source and microphones within these simulated environments under the following constraints. The sound source was placed at a fixed height of 1.75 m, with its $x$ and $y$ coordinates randomly determined within the room, ensuring a minimum distance of 0.5 m from the room walls. Similarly, the microphones were positioned at a fixed height of 1.6 m, with their $x$ and $y$ coordinates also randomly determined within the room dimensions. The reverberation time was set at the range

[2] `github.com/microsoft/Swin-Transformer`

[3] `github.com/DavidDiazGuerra/gpuRIR`

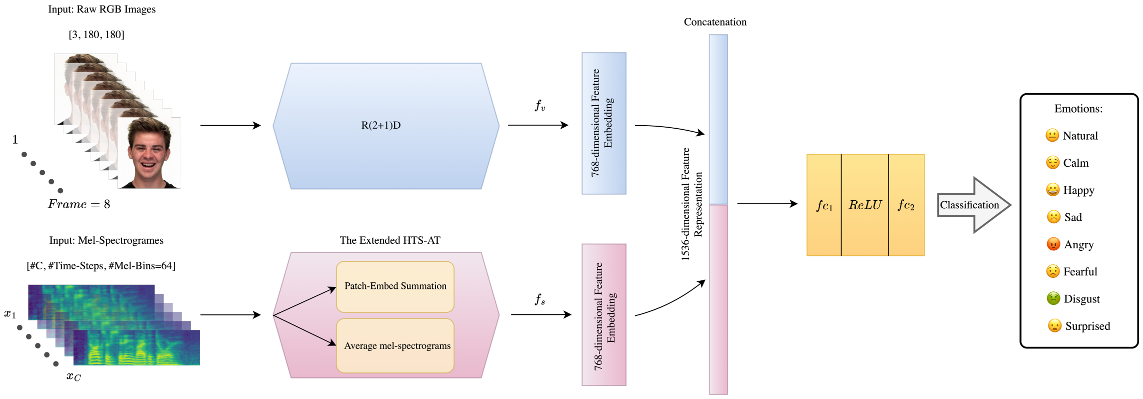


Fig. 1: The proposed Multi-modal Emotion Recognition (MER).

$T_{60} = 500 - 850$ ms. The distance between the sound source and microphones was randomly selected in the range $[0.2, d_c]$ m, where $d_c$ to the critical distance, determined by the room's volume and $T_{60}$. This ensures the dominance of direct sound over reflections. Finally, we added spatially-white noise with signal-to-noise ratio (SNR) of 20 dB to each reverberant signal. This noise was synthesized by applying an auto-regressive filter of order 1 to a white Gaussian noise, emphasizing lower frequencies.

The proposed scheme was evaluated using real-world RIRs drawn from the ACE database [21]. The ACE RIR database comprises recordings from seven different rooms with varying dimensions and reverberation levels (see Table II). We only used a subset of the database, recorded with a mobile phone equipped with three microphones in the near-field scenario, which is a practical choice for real-world SER applications. We convolved all audio utterances of the test set with the ACE RIRs to generate 3-microphone signals for each utterance.

### B. Algorithm Setup

As discussed earlier, the video modality leverages the $R(2+1)$D model pre-trained on the action recognition Kinetics dataset [26]. To better suit our ER task, we modified the model's architecture by adjusting the final linear layer. Specifically, we reconfigured it to output 768-dimensional feature embeddings. This adjustment ensures that both modalities (video and audio) contribute equally-sized feature vectors to the multi-modal representation by fusion through concatenation.

The resolution of the RGB video frames was first reduced to $224 \times 224$ pixels. Then, eight frames from the video stream were randomly selected. To augment the dataset, these frames underwent refinement through random cropping using the TorchVision Library [23], yielding $180 \times 180$ images that enhance the model's robustness to spatial variations. In addition, we added random horizontal and vertical flips, each with a 30% probability of application, coupled with arbitrary rotations within the range of [-30°, 30°].

The audio modality applies an extended version of the HTS-AT model [18], suitable for both multi-channel and single-channel scenarios. As described in Sec. III, the network structure configuration is arranged into four groups, each containing several Swin-Transformer blocks: 2, 2, 6, and 2, respectively. The mel-spectrogram input is initially transformed into patches and linearly projected to a dimension of $D = 96$. This dimension expands exponentially through each transformer group, finally reaching a dimension of 768 ($8D = 768$), which matches the design principles of Audio Spectrogram Transformer (AST). Pre-processing was carried out as explained in [18] both for multi-channel and single-channel experiments. We augmented the mel-spectrograms by using the SpecAugment Library [22], which consists of temporal masking, occluding four distinct "strips", each 64 time-frames long. Complementing this, we applied frequency domain masking, obscuring two strips, each 8 frequency bands wide.

Our multi-modal approach combines the feature embeddings from both the video and audio modalities. The 768-dimensional feature vectors extracted from the $R(2+1)$D model and the extended HTS-AT model are concatenated, resulting in a 1536-dimensional feature representation. This combined feature vector is then fed into a classification head for prediction. The right-hand side of Fig. 1 presents two fully connected layers $(fc)$ with a Relu activation function between them, forming the sequence:

$$f_{c_1} \rightarrow \text{ReLU} \rightarrow f_{c_2} \;\Rightarrow E \tag{4}$$

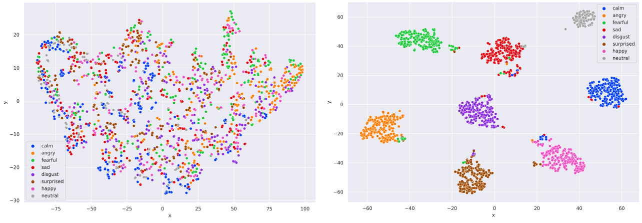
The fine-tuning processes are applied using the Adam optimizer with a learning rate of $1e^{-3}$ and a warm-up strategy. We used cross-entropy loss as the metric with a batch size of 32. The maximum number of epochs was set to 500 for all experiments, with an early stopping strategy with a patience of 12 to prevent overfitting. In practice, the maximum number of epochs was never reached, as the fine-tuning process was halted earlier due to the activation of the patience parameter. The overall number of parameters for the fine-tuned models are as follows: 32.3M for the uni-modal scheme based on video, 28.7M for the uni-modal scheme based on audio, and 62.7M for the multi-modal scheme.

TABLE I: The Accuracy results of single-microphone MER method compared with SOTA MER methods tested on the original RAVDESS dataset. Results for the competing methods are taken from the corresponding articles.

| **Methods** | **ACC (%)** |
|---|---|
| Human performance [20] | 80.00 |
| Garaiman et al. [27] | 65.76 |
| Ghaleb et al. [28] | 76.30 |
| Franceschini et al. [8] | 78.54 |
| Radoi et al. [29] | 78.70 |
| **Luna-Jiménez et al.** [30] | **80.08** |
| Proposed MER ($C = 1$) | 80.00 |

### C. Results

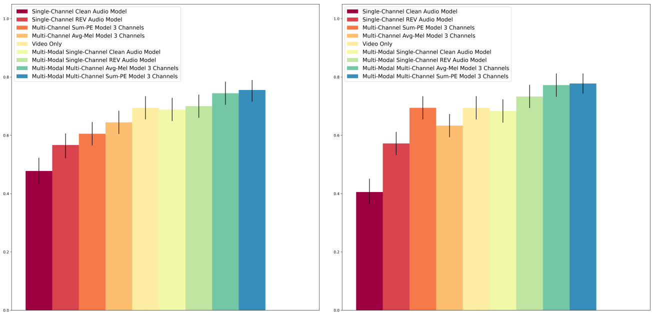
Table I compares the performance of the proposed MER approach (single-microphone variant, $C = 1$) with several state-of-the-art (SOTA) MER approaches evaluated on RAVDESS. The results indicate that our single-microphone MER achieves performance on par with [8], [29], [30]. Moreover, to assess and visualize the separation capabilities of the proposed scheme across the clean RAVDESS dataset, we employed the t-distributed Stochastic Neighbor Embedding (t-SNE) visualization method. This nonlinear technique reduces high-dimensional data into two- or three-dimensional representations suitable for graphical visualization. Importantly, it maps nearby points in the high-dimensional space to close points in the reduced space, while far-apart points remain distant in the visualization [31]. In Fig. 2, we compare the t-SNE mapping of the input features extracted from the RAVDESS dataset and the network output, depicting each emotion with a unique color and shape to visualize the clustering quality. The enhancement in classification performance following the network's application is immediately apparent.



(a) RAVDESS: Features (b) RAVDESS: Network output

Fig. 2: t-SNE visualization.

We now turn to the evaluation of the multi-channel schemes, using the reverberant RAVDESS dataset version, applying MER with a multi-microphone ($C = 3$). Table II details our Accuracy results of the various emotion recognition schemes for seven different rooms from the ACE database. The video-only modality is compared with the audio-only modality (both single- and multi-channel models) and the combined multi-modal approach. As the video modality is unaffected by the acoustic conditions, we only report the results once. We investigated two single-channel ($C = 1$) variants: one fine-tuned on clean speech and the other on reverberant speech. Both were evaluated using a single microphone from the ACE test set. To assess the reliability of our results, we report the mean results together with 75% confidence intervals.[4]



(a) RAVDESS in ACE Office 2 ($T_{60} = 390$ ms). (b) RAVDESS in ACE Lobby ($T_{60} = 646$ ms).

Fig. 3: Accuracy and confidence intervals assessed using the reverberated RAVDESS test set for two different rooms from the ACE database.

Analyzing Table II, it is observed that for the audio-only schemes, the multi-channel processing methods (Avg mel and Sum PE) consistently outperform the single-channel approaches. This is in line with the findings of [18]. Notably, the multi-modal approaches significantly outperform their uni-modal counterparts. These results are also visually demonstrated in Fig. 3, demonstrating the advantages of multi-modal processing. In addition, Fig. 4 presents the confusion matrix for the multi-channel Sum PE MER model. The confusion matrix compares the actual target and predicted labels, showing the percentage of correct and incorrect predictions for each class. Beyond measuring accuracy, it also reveals the distribution of errors across different emotions, helping to identify specific misclassifications.

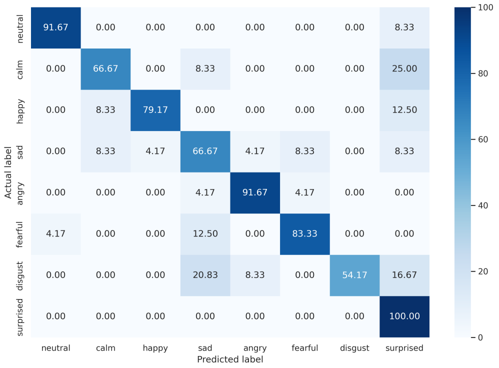


Fig. 4: Confusion matrix of the results of multi-channel Sum PE MER model on RAVDESS test set convolved with ACE Lecture Room 2 ($T_{60} = 1220$ ms).

## V. Conclusions

In this paper, we presented a MER system designed to operate in reverberant and noisy acoustic environments. Our approach demonstrates robust performance across a range of

[4] github.com/luferrer/ConfidenceIntervals

TABLE II: Accuracy and the associated confidence intervals of the proposed method for the RAVDESS test set reverberated by RIRs drawn from the ACE database (using the 3-microphone of the near-filed cellular phone). The 'Single-Channel' columns use an arbitrarily chosen microphone, fine-tuned on either clean or reverberant data, respectively. The 'Avg mel' columns present results with mel-spectrograms averaged across three channels during fine-tuning and testing. The 'Sum PE' columns depict the Patch-Embed fusion approach fine-tuned and tested on the three channels. The asterisk in the video column describes the same result. The best results for each modality are underlined, while the overall best result is shown in boldface.

| Room ($T_{60}$ [ms]) | Video | Audio | | | | MER | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Clean Single-Channel | Rev Single-Channel | Avg mel | Sum PE | Clean Single-Channel | Rev Single-Channel | Avg mel | Sum PE |
| Lecture 1 (638) | 69.4 (65.5-73.3) | 42.7 (38.8-47.2) | 60 (55.5-64.4) | 61.6 (57.7-65.5) | 61.1 (57.2-65) | 70 (66.1-73.8) | 75.0 (71.1-78.8) | 77.2 (73.3-81.1) | **78.3** (74.4-81.6) |
| Lecture 2 (1220) | * | 40.5 (36.6-45) | 55 (50.5-59.4) | 60.5 (56.1-64.4) | 58.8 (54.9-62.8) | 71.6 (67.7-75.5) | 76.1 (72.7-79.4) | 76.1 (72.7-80) | **78.3** (75.0-81.6) |
| Lobby (646) | * | 40.5 (36.6-45) | 57.2 (53.3-61.1) | 63.3 (59.4-67.22) | 69.4 (65.5-73.3) | 68.3 (64.4-72.2) | 73.3 (69.4-77.2) | 77.2 (73.3-81.1) | **77.7** (74.4-81.1) |
| Meeting 1 (437) | * | 42.7 (38.8-46.6) | 57.2 (52.7-61.6) | 60 (56.1-63.8) | 61.1 (57.2-65) | 70 (66.1-73.8) | 72.7 (68.8-76.6) | 77.2 (73.8-81.1) | **78.8** (75.5-82.2) |
| Meeting 2 (371) | * | 38.8 (34.4-42.7) | 56.1 (51.6-60.5) | 62.7 (58.8-66.6) | 59.4 (55.5-63.8) | 71.1 (67.2-75) | 75 (71.1-78.8) | **78.8** (75.5-82.2) | 75.5 (71.6-78.8) |
| Office 1 (332) | * | 42.7 (38.3-46.6) | 59.4 (55-63.3) | 62.7 (58.8-68.3) | 63.3 (58.8-67.2) | 70 (66.1-73.8) | 75 (71.1-78.8) | **77.7** (74.4-81.6) | 77.2 (73.8-80.5) |
| Office 2 (390) | * | 47.7 (43.3-52.2) | 56.6 (52.2-60.5) | 64.4 (60.5-68.3) | 60.5 (56.6-64.4) | 68.8 (65-72.7) | 70 (66.1-73.8) | 74.4 (70.5-78.3) | **75.5** (71.6-78.8) |

realistic acoustic conditions by combining an extended multi-channel HTS-AT for audio processing with an $R(2+1)$D model for video analysis. The MER system combines audio and visual modalities, consistently outperforming uni-modal approaches. Using synthetic RIRs for training and real-world RIRs from the ACE database for testing, we comprehensively assess our system's performance in diverse acoustic environments. Moreover, the utilization of multi-channel audio processing, particularly the Patch-Embed summation, proves beneficial in mitigating the effects of reverberation and noise over the single-channel case. This leads to the potential of our MER system for applications in various real-world scenarios where acoustic conditions may be far from ideal. Future work could further improve the system's performance in extremely reverberant environments and explore its effectiveness in other emotional datasets.

## References

[1] S. K. Bharti *et al.*, "Text-based emotion recognition using deep learning approach," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 2645381, 2022.

[2] D. Sherman, G. Hazan, and S. Gannot, "Study of speech emotion recognition using BLSTM with attention," in *European Signal Processing Conf. (EUSIPCO)*, Helsinki, Finland, Sep. 2023.

[3] S. Zhou *et al.*, "Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks," *Int. Journal of Environmental Research and Public Health*, vol. 20, no. 2, p. 1400, 2023.

[4] V. John and Y. Kawanishi, "Audio and video-based emotion recognition using multimodal transformers," in *Int. Conf. on Pattern Recognition (ICPR)*, 2022, pp. 2582–2588.

[5] F. Noroozi *et al.*, "Audio-visual emotion recognition in video clips," *IEEE Trans. on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2017.

[6] W. Dai *et al.*, "Multimodal end-to-end sparse model for emotion recognition," *arXiv preprint arXiv:2103.09666*, 2021.

[7] I. Kansizoglou, L. Bampis, and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Trans. on Affective Computing*, 2019.

[8] R. Franceschini *et al.*, "Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss," in *International Conference on Pattern Recognition (ICPR)*, 2022, pp. 2589–2596.

[9] C.-W. Huang and S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2017, pp. 583–588.

[10] M. Seo and M. Kim, "Fusing visual attention cnn and bag of visual words for cross-corpus speech emotion recognition," *Sensors*, vol. 20, no. 19, p. 5559, 2020.

[11] Y. Song, Y. Cai, and L. Tan, "Video-audio emotion recognition based on feature fusion deep learning method," in *IEEE Int. Midwest Symposium on Circuits and Systems (MWSCAS)*, 2021, pp. 611–616.

[12] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 552–558.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conf. on computer vision and pattern recognition*, 2015, pp. 815–823.

[14] A. B. Zadeh *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*, 2018, pp. 2236–2246.

[15] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[16] T. Mittal *et al.*, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *AAAI Conf. on Artificial Intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[17] J.-B. Delbrouck, N. Tits, and S. Dupont, "Modulated fusion using transformer for linguistic-acoustic emotion recognition," *arXiv preprint arXiv:2010.02057*, 2020.

[18] O. Cohen, G. Hazan, and S. Gannot, "Multi-microphone speech emotion recognition using the hierarchical token-semantic audio transformer architecture," *arXiv preprint arXiv:2406.03272*, 2024, accepted to IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2025.

[19] D. Tran *et al.*, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[20] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[21] J. Eaton *et al.*, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[22] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019, pp. 2613–2617.

[23] T. maintainers and contributors, "TorchVision: PyTorch's Computer Vision library," `https://github.com/pytorch/vision`, 2016.

[24] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF Int. Conf. on computer vision*, 2021, pp. 10 012–10 022.

[25] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. on Learning Representations (ICLR)*, 2021.

[26] W. Kay *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[27] F. E. Garaiman and A. Radoi, "Multimodal emotion recognition system based on x-vector embeddings and convolutional neural networks," in *International Conference on Communications (COMM)*, 2024.

[28] E. Ghaleb, J. Niehues, and S. Asteriadis, "Multimodal attention-mechanism for temporal emotion recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 251–255.

[29] A. Radoi *et al.*, "An end-to-end emotion recognition framework based on temporal aggregation of multimodal information," *IEEE Access*, vol. 9, pp. 135 559–135 570, 2021.

[30] C. Luna-Jiménez *et al.*, "Multimodal emotion recognition on RAVDESS dataset using transfer learning," *Sensors*, vol. 21, no. 22, p. 7665, 2021.

[31] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.