Leveraging Facial and Vocal Cues for Emotion Recognition: A Transformer-Based Approach

Abstract

Emotion recognition is a cornerstone of affective computing, enabling systems to understand and respond to human emotions effectively. Multimodal emotion recognition, combining facial and vocal cues, offers a more robust understanding of affective states due to the intricate interplay between these modalities. Using the RAVDESS dataset, our project explores three multimodal emotion recognition frameworks including two transformer-based approaches and one attention-based approach to incorporate both facial and vocal cues. We concentrate on the strengths of both facial and vocal signals separately and in combination, whereas traditional models frequently treat them as separate data streams. While vocal intonations capture more subtle, auditory cues of affect, facial expressions convey visible emotional nuances. We obtain a reliable and scalable solution by using decision-level fusion and transformer models for every modality. Our best performing approach,Intermediate Attention Fusion, achieves 33.958% top-1 precision and 98.125% top-5 precision on the RAVDESS dataset. In addition to increasing accuracy in multimodal contexts, this work shows how each modality contributes differently to emotion recognition. This project hopes to contribute to advancing emotion recognition in applications such as sentiment analysis, human-computer interaction, and adaptive user interfaces.

1. Introduction

Facial expressions are fundamental to human communication, conveying emotions and mental states both consciously and unconsciously throughout our daily interactions. These visual signals provide insights into a person's emotional state and thought processes. The automatic recognition of facial expressions has numerous practical applications across different fields. In healthcare, it can assist in early detection of neurological conditions like Parkinson's disease, where facial expressiveness often diminishes, or help monitor patients' pain levels. In automotive safety, systems can detect driver fatigue or distraction through facial cues. In professional settings, facial expression analysis can enhance remote interviews and customer service interactions by providing feedback on engagement and emotional responses. Law enforcement agencies may use this technology to support interrogation processes, though with careful consideration of accuracy limitations.

Recent research has made significant progress in multimodal emotion recognition. While some of systems have shown promising results, our research aims to advance multimodal emotion recognition by incorporating transformer architectures for both facial and speech analysis, leveraging their proven capability to model complex temporal dependencies across modalities. Our research explores three different approaches for combining facial and vocal cues: Late Transformer Fusion, Intermediate Transformer Fusion, and Intermediate Attention-Based Fusion. Our method processes each modality independently through separate branches, then combines them using these different fusion strategies. We compare these approaches to understand their effectiveness in multimodal emotion recognition using the RAVDESS dataset.

2. Related Work

Emotion recognition has been a significant area of research, studied through different modalities such as speech (Speech Emotion Recognition, SER), facial expressions (Facial Emotion Recognition, FER), and their combinations. Researchers have explored a variety of machine learning and deep learning techniques to enhance performance across these domains.

Kumar et al. [4] demonstrated successful emotion recognition on RAVDESS using transfer learning techniques, achieving high accuracy through multimodal fusion. Chen et al. [2] proposed EmotiCon, which leverages Frege's Principle to combine facial, vocal and contextual cues for enhanced recognition performance.

Pandey et al. [8] presented a comprehensive review of deep learning-based approaches for SER. Their study examined features such as spectrograms, Mel spectrograms, and Mel-frequency cepstral coefficients (MFCCs), emphasizing the effectiveness of hybrid CNN and LSTM models

in capturing both spatial and temporal emotional features. Yue et al. [15] proposed a multi-task learning framework that leverages Wav2Vec 2.0 for self-supervised speech representations. This framework simultaneously predicts emotion categories and intensity levels, demonstrating the utility of pre-trained models in SER tasks. Xia and Liu [14] implemented a Deep Belief Network (DBN) for feature extraction, integrating multi-task learning to predict both categorical emotions and continuous attributes like valence and arousal.

Advanced methods combining neural architectures and attention mechanisms have also been proposed. Zhang et al. [17] developed an AlexNet-inspired Fully Convolutional Network (FCN) with an integrated attention mechanism. By focusing on critical time-frequency regions of spectrograms, their model significantly improved SER accuracy. Sherman et al. [11] introduced a hybrid CLDNN model that merges CNNs, BiLSTMs, and attention mechanisms to capture intricate temporal patterns in speech signals. Shah et al. [10] focused on multimodal approaches, integrating prosodic, spectral, and facial features. They employed a linear Support Vector Machine (SVM) for decision-level fusion, which enhanced overall performance.

Multimodal fusion has been a growing area of focus in emotion recognition research. Cai et al. [1] proposed a framework combining CNNs and LSTMs for both SER and FER, where outputs were fused using deep neural networks. Patamia et al. [9] employed transformers for SER by utilizing BERT embeddings for textual features and neural networks for motion capture (MoCap) data, achieving fusion at the prediction stage. Tripathi et al. [12] used BiLSTMs with attention mechanisms to process spectral features for SER, alongside MoCap data for FER, demonstrating the advantages of combining features across modalities.

These studies highlight the progress in emotion recognition, emphasizing the benefits of hybrid architectures, attention mechanisms, and multimodal fusion techniques. Together, these advancements continue to drive innovations in the field, offering more accurate and robust emotion classification systems.

3. Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [6] is a comprehensive multimodal dataset designed for emotion recognition research5. It contains 7,356 files featuring 24 professional actors (12 female, 12 male) vocalizing two lexically-matched statements in a neutral North American accent. The dataset includes both speech and song recordings, covering a range of emotions: calm, happy, sad, angry, fearful, surprise, and disgust. Each expression is produced at two levels of emotional intensity (normal and strong), with an additional neutral expression. RAVDESS offers three modality formats: Audio-only (16bit, 48kHz .wav files), Audio-Video (720p H.264, AAC 48kHz, .mp4 files), and Video-only (no sound).

Figure 1. RAVDESS Audio_Song_Actors_01-24

This diverse set of modalities allows researchers to explore unimodal and multimodal approaches to emotion recognition. The dataset's balanced gender representation and professional acting ensure high-quality, consistent emotional expressions. RAVDESS has been validated through extensive rating studies, with each recording rated 10 times on emotional validity, intensity, and genuineness by 247 untrained individuals from North America5. The dataset's structured nature, with controlled lexical content and emotional intensities, makes it particularly suitable for developing and evaluating multimodal emotion recognition systems. Its inclusion of both speech and song also opens up possibilities for comparing emotion recognition across different vocal modalities.

The RAVDESS dataset offers several key advantages:

- **Equal Representation:** Equal gender distribution (12 Male, 12 Female), ensures unbiased analysis across the dataset.
- **Lexical Uniformity:** Identical speech content across samples minimizes linguistic bias, enhancing the reliability of the dataset.
- **Validation:** Each recording has been independently rated by a diverse pool of participants, ensuring reliable emotion labeling and high-quality data.

4. Dataset Preprocessing

To guarantee compatibility with transformer architectures and optimize model performance, the RAVDESS data underwent the following preprocessing steps in our project:

Processing Audio

- Resampled stereo recordings to 16 kHz and converted them to mono.

- Focused on pitch and tone variations that reflect emotional intonations by extracting Mel-Frequency Cepstral Coefficients (MFCCs) as audio features.
- Applied amplitude normalization to reduce variability across recordings.

Processing Video

- Used OpenFace to extract facial regions and resized them to 224 × 224 pixels.
- Uniformly selected 15 frames from each video to capture temporal emotional dynamics while minimizing computational overhead.
- Applied data augmentation techniques, including rotations and horizontal flips, to improve generalizability.

Syncing

- Temporally aligned audio and visual data to maintain emotional coherence between modalities.
- Padded or truncated sequences to uniform lengths to address variable durations in the dataset.

Normalization

- Standardized feature values to zero-mean and unit-variance distributions, ensuring stability during training.

5. Method

This project focuses on audiovisual emotion recognition using an end-to-end trainable model to overcome challenges posed by missing or noisy modalities. The architecture integrates separate branches for audio and visual feature extraction, along with transformer-based modality fusion strategies [13].

5.1. Feature Extraction

5.1.1 Vision Branch

The vision branch processes raw video sequences through a two-stage pipeline. Initially, individual video frames are passed through the EfficientFace model, pre-trained on the AffectNet dataset [7], to extract meaningful features. This step avoids reliance on pre-extracted features like facial landmarks or action units, facilitating an end-to-end learning approach. The extracted features are then processed using temporal 1D convolutional blocks to capture temporal dependencies across frames. Specifically, each block consists of a 1D convolutional layer with a 3x3 kernel, followed by batch normalization, ReLU activation, and grouping into two stages. This design balances computational efficiency and the need for temporal representation in emotion recognition tasks [16].

5.1.2 Audio Branch

The audio branch takes mel-frequency cepstral coefficients (MFCCs) as input, a widely used feature set in speech analysis. Four convolutional blocks, each comprising a convolutional layer, batch normalization, ReLU activation, and max-pooling, are applied to learn hierarchical representations. These blocks progressively capture both low-level and high-level features essential for distinguishing emotional states. Unlike alternate representations such as chroma features or spectrograms, MFCCs were found to yield optimal performance in this context.

Table 1. Architecture of the Visual and Audio Branches

Visual Branch (EfficientFace Module)	
Stage 1	Input Reshaping Conv1D [Kernel: 3, Output: 64, Stride: 1] + BatchNorm1D + ReLU Conv1D [Kernel: 3, Output: 64, Stride: 1] + BatchNorm1D + ReLU
Stage 2	Conv1D [Kernel: 3, Output: 128, Stride: 1] + BatchNorm1D + ReLU Conv1D [Kernel: 3, Output: 128, Stride: 1] + BatchNorm1D + ReLU
Prediction	Global Average Pooling + Linear Layer
Audio Branch	
Stage 1	Conv1D [Kernel: 3, Output: 64] + BatchNorm1D + ReLU + MaxPooling [2x1] Conv1D [Kernel: 3, Output: 128] + BatchNorm1D + ReLU + MaxPooling [2x1]
Stage 2	Conv1D [Kernel: 3, Output: 256] + BatchNorm1D + ReLU + MaxPooling [Kernel: 2] Conv1D [Kernel: 3, Output: 128] + BatchNorm1D + ReLU + MaxPooling [Kernel: 2]
Prediction	Global Average Pooling + Linear Layer

5.2. Modality Fusion Approaches

To effectively integrate audio and visual information, three distinct fusion mechanisms were explored: Late Transformer Fusion, Intermediate Transformer Fusion, and Intermediate Attention-Based Fusion. These approaches are designed to combine features from the two modalities [5] while addressing potential co-dependencies and ensuring robustness in incomplete data scenarios.

5.2.1 Late Transformer Fusion

In this approach, the features from the audio and visual branches are processed independently through their respective convolutional layers. Fusion occurs at the final stage of the pipeline using transformer blocks. Specifically, a transformer block is employed for each modality, enabling cross-attention between the two branches.

For example, let $\Phi_a \in \mathbb{R}^{T \times d_a}$ and $\Phi_v \in \mathbb{R}^{T \times d_v}$ represent the feature matrices of the audio and visual modalities, respectively, after their final convolutional layers. The audio branch computes queries $\mathbf{Q}_a = \Phi_a W_q$ and attends to keys and values derived from the visual branch, $\mathbf{K}_v = \Phi_v W_k$ and $\mathbf{V}_v = \Phi_v W_v$. The attention output for the audio branch is then given by:

$$\mathbf{A}_a = \text{softmax}\left(\frac{\mathbf{Q}_a \mathbf{K}_v^\top}{\sqrt{d_k}}\right)\mathbf{V}_v,$$

where d_k is the dimensionality of the key vectors, and W_q, W_k, and W_v are learnable weight matrices. A similar computation is performed for the visual branch. The outputs of

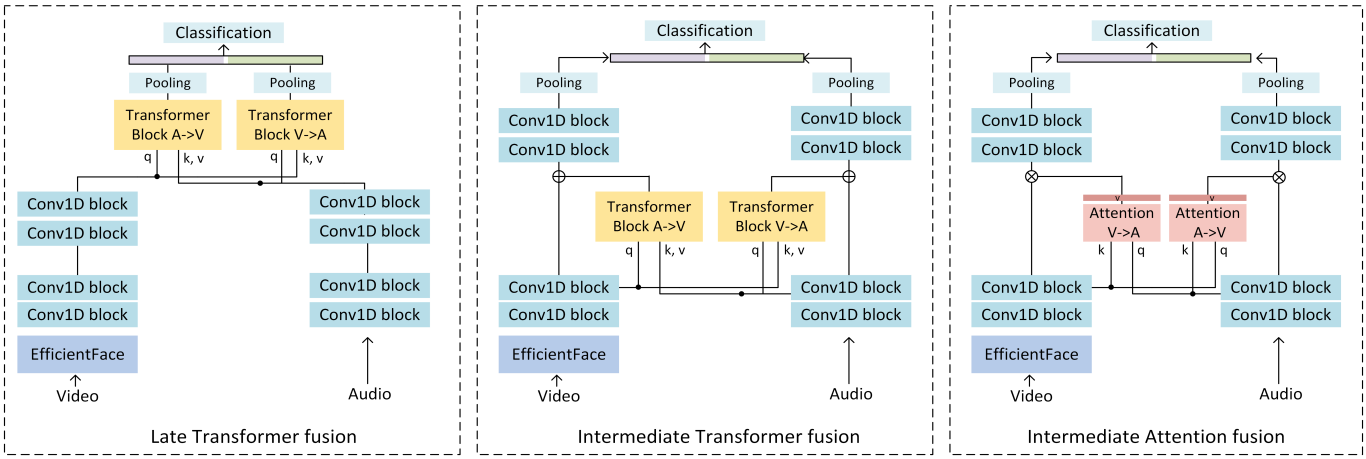

Figure 2. Modality Fusion Approaches

the transformer blocks are concatenated and passed to the final classification layer.

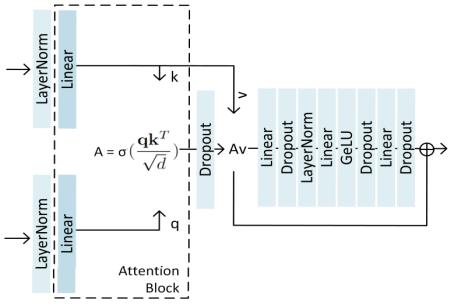

Figure 3. Transformer Block

5.2.2 Intermediate Transformer Fusion

In this approach, modality fusion is performed at intermediate feature layers, enabling early interaction between audio and visual representations. After the first stage of feature extraction (e.g., after two convolutional layers in each branch), transformer blocks are applied for cross-modal attention. The outputs are added back to the respective branches, allowing joint feature learning in subsequent layers.

Formally, let $\Phi_a^{(1)}$ and $\Phi_v^{(1)}$ denote the intermediate features from the audio and visual branches. Cross-modal attention is computed using the same mechanism as in Late Transformer Fusion:

$$\mathbf{A}_a^{(1)} = \text{softmax}\left(\frac{\mathbf{Q}_a^{(1)}\mathbf{K}_v^{(1)\top}}{\sqrt{d_k}}\right)\mathbf{V}_v^{(1)},$$

where $\mathbf{Q}_a^{(1)}$, $\mathbf{K}_v^{(1)}$, and $\mathbf{V}_v^{(1)}$ are projections of $\Phi_a^{(1)}$ and $\Phi_v^{(1)}$ through the corresponding weight matrices. The resulting attention outputs are added to the original features:

$$\Phi_a^{(1)} \leftarrow \Phi_a^{(1)} + \mathbf{A}_a^{(1)}, \quad \Phi_v^{(1)} \leftarrow \Phi_v^{(1)} + \mathbf{A}_v^{(1)}.$$

The modified features are passed through the second stage of feature extraction, allowing the network to refine its joint representations.

5.2.3 Intermediate Attention-Based Fusion

This systematic approach uses a simplified attention mechanism to identify relevant features across modalities without enforcing direct co-dependencies [3]. Unlike the transformer-based methods, the attention mechanism operates on scaled dot-product similarity alone.

Given intermediate features $\Phi_a^{(1)} \in \mathbb{R}^{T\times d_a}$ and $\Phi_v^{(1)} \in \mathbb{R}^{T\times d_v}$, the attention matrix $\mathbf{A} \in \mathbb{R}^{T\times T}$ is computed as:

$$\mathbf{A} = \text{softmax}\left(\frac{\Phi_a^{(1)} W_q W_k^\top \Phi_v^{(1)\top}}{\sqrt{d_k}}\right).$$

The attention scores are used to highlight the most relevant attributes of each modality. For example, the attention-modulated features of the audio branch are computed as:

$$\mathbf{v}_a = \sum_{t=1}^{T} \mathbf{A}[:, t],$$

where $\mathbf{A}[:, t]$ represents the t-th column of the attention matrix. This vector aggregates the relative importance of audio features in the context of visual features. A similar computation is performed for the visual branch. Unlike traditional attention mechanisms, this method avoids direct feature fusion, focusing instead on shared information while reducing over-reliance on one modality.

5.2.4 Comparison and Benefits

- **Late Transformer Fusion:** Maximizes the individual learning capacity of each branch but risks modality-specific overfitting.
- **Intermediate Transformer Fusion:** Balances early integration and modality-specific learning, enhancing joint representation capabilities.
- **Intermediate Attention-Based Fusion:** Focuses on complementary features, improving robustness to missing or noisy modalities.

These approaches ensure flexibility and adaptability, catering to various real-world challenges, including incomplete or noisy data scenarios.

6. Experiments and Results

For the experiments, we run the three approaches mentioned above–Late Transformer, Intermediate Transformer Fusion, and Intermediate Attention Fusion on the preprocessed RAVDESS dataset and compare the predicted outputs with ground truth emotions, with the following configuration:

6.1. Training Protocol

The model was trained end-to-end on the RAVDESS dataset, using EfficientFace for visual features and MFCCs for audio features. Data augmentation techniques, such as random horizontal flips and rotations, were applied to improve generalization. Training was performed using Stochastic Gradient Descent (SGD) with a learning rate of 0.04, a momentum of 0.9, and a weight decay of 10^{-3}. The learning rate was reduced on a plateau after 10 epochs. The model was trained for 100 epochs before evaluations.

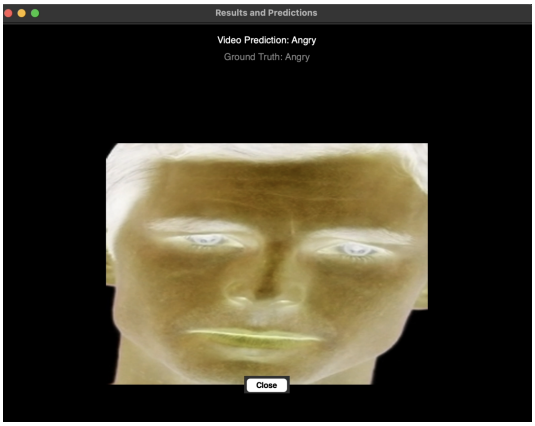

Figure 4. Predicted Result on RAVDESS dataset using IA: Angry

6.2. Hyperparameter Selection

Our hyperparameter choices were determined through ablation studies:

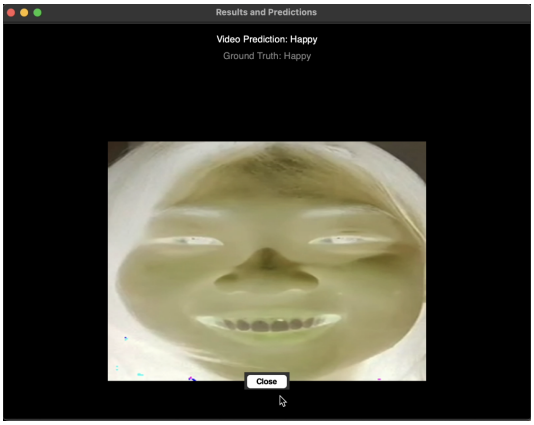

Figure 5. Predicted Result on custom dataset using IA: Happy

6.2.1 Learning Rate (0.04)

- Tested range [0.01, 0.04, 0.1]
- 0.04 provided fastest convergence without stability issues
- Higher rates led to training instability
- Lower rates resulted in slower convergence

6.2.2 Weight Decay (10^{-3})

- Tested range [10^{-4}, 10^{-3}, 10^{-2}]
- 10^{-3} provided optimal regularization without underfitting

6.3. Evaluation and Results

We evaluated three appoaches using:

loss: the training loss value

prec1: top-1 precision (percentage of times the model's top prediction was correct)

prec5: top-5 precision (percentage of times the correct emotion was among the model's top 5 predictions)

custom test dataset: custom dataset of 12 new videos featuring individuals expressing different emotions. This custom dataset serves as an independent test set to validate the real-world applicability of our approaches

	loss	prec1	prec5
Late Transformer	16.699	14.375	59.167
Intermediate Transformer Fusion	35.392	13.958	85.208
Intermediate Attention Fusion	2.393	33.958	98.125

Table 2. Experiment Results

We achieved 66.7% accuracy on the custom dataset using the best performing approach– Intermediate Attention. While this performance suggests reasonable generalization,

the limited emotion range and small sample size prevent definitive conclusions about real-world performance.

6.4. Analysis

Experiments were conducted on three fusion approaches and achieved different results as shown in the above table.

The Late Transformer approach achieved a moderate loss value of 16.699, with precision metrics showing 14.375% for prec1 and 59.167% for prec5. This suggests that while the model performs reasonably well when considering its top-5 predictions, its top-1 accuracy is relatively low, indicating some uncertainty in its primary predictions.

The Intermediate Transformer Fusion has the highest loss value at 35.392, more than double that of the Late Transformer approach. However, interestingly, it achieved a markedly better prec5 score of 85.208%, despite a slightly lower prec1 of 13.958%. This indicates that while the model may struggle with precise single-class predictions, it demonstrates strong performance in identifying the correct emotion within its top-5 predictions.

The Intermediate Attention Fusion approach results in the best performance across three approaches. With the lowest loss value of 2.393, it significantly outperformed both transformer-based approaches in terms of training convergence. More importantly, it achieved the highest precision scores, with a prec1 of 33.958% and an impressive prec5 of 98.125%. This improvement in both metrics suggests that the simplified attention mechanism's focus on only the most important features from face and voice separately, rather than enforcing the model to find relationships between every face and voice feature, eventually results in more robust and accurate emotion recognition.

These results demonstrate a clear trade-off between architectural complexity and performance. While transformer-based approaches offer sophisticated mechanisms for modal integration, the simpler attention-based fusion proves more effective in practice. This suggests that for multimodal emotion recognition tasks, picking out the most important features from each input type may be more effective than complex processing.

References

[1] Linqin Cai, Jiangong Dong, and Min Wei. Multi-modal emotion recognition from speech and facial expression based on deep learning. In *2020 Chinese Automation Congress (CAC)*, pages 5726–5729, 2020.

[2] Bin Chen, Lei Wang, and Min Zhang. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1245, 2022.

[3] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Self-attention fusion for audiovisual emotion recognition with incomplete data, 2022.

[4] Ashish Kumar, Rajesh Sharma, and Dhruba Bhattacharyya. Multimodal emotion recognition on ravdess dataset using transfer learning. *IEEE Access*, 9:38400–38417, 2021.

[5] Okan Köpüklü, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1910–1919, 2019.

[6] Steven R Livingstone and Frank 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*, 13(5): e0196391, 2018.

[7] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *CoRR*, abs/1708.03985, 2017.

[8] Sandeep Pandey, H Shekhawat, and S Prasanna. Deep learning techniques for speech emotion recognition : A review. 2019.

[9] Rutherford Agbeshi Patamia, Wu Jin, Kingsley Nketia Acheampong, Kwabena Sarpong, and Edwin Kwadwo Tenagyei. Transformer based multimodal speech emotion recognition with improved neural networks. In *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 195–203, 2021.

[10] Mohit Shah, Chaitali Chakrabarti, and Andreas Spanias. A multi-modal approach to emotion recognition using undirected topic models. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 754–757, 2014.

[11] Dalia Sherman, Gershon Hazan, and Sharon Gannot. Study of speech emotion recognition using blstm with attention. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 416–420, 2023.

[12] Samarth Tripathi and Homayoon S. M. Beigi. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *CoRR*, abs/1804.05788, 2018.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[14] Rui Xia and Yang Liu. Leveraging valence and activation information via multi-task learning for categorical emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5301–5305, 2015.

[15] Pengcheng Yue, Leyuan Qu, Shukai Zheng, and Taihao Li. Multi-task learning for speech emotion and emotion intensity recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1232–1237, 2022.

[16] Peng Zhang, Feng Zhao, Peng Liu, and Mengwei Li. Efficient lightweight attention network for face recognition. *IEEE Access*, 10:31740–31750, 2022.

[17] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1775, 2018.