A Multimodal Framework for Emotion Recognition

Abstract—**Emotion recognition is critical in many fields, from customer services to human-computer interaction. This study proposes a multimodal framework for emotion recognition, which integrates images and audio, Using pre-trained models specifically designed for emotion recognition tasks, such as Emotion2Vec and Vision Transformer (ViT) for feature extraction. Our approach captures and combines features across modalities. In the benchmark of the RAVDESS dataset, our multimodal method showed robust and effective performance, achieving results comparable to the state-of-the-art. Although the average accuracy did not exceed the state-of-the-art benchmark, our model produced competitive results and, by one fold, even exceeded the reported accuracy, highlighting its potential in specific scenarios. However, an attempt to validate the model on a custom-created dataset was unsuccessful, highlighting areas for further improvement. These findings emphasize the potential of multimodal systems in advancing emotion recognition technologies while underscoring the importance of robust dataset generalization.**

Index Terms—**Multimodal Framework, Emotion Recognition, RAVDESS Dataset, Pre-trained Models, Feature Extraction, Transfer Learning, Emotion2Vec, Vision Transformer (ViT), Speech Emotion Recognition (SER), Facial Emotion Recognition (FER)**

I. Introduction

Emotions are essential in shaping human interactions, significantly affecting communication in physical and digital environments. Observing and responding to emotional signs in online meetings, customer service platforms, and digital environments is important for gaining trust, improving user experiences, and driving better outcomes. Emotion recognition has advanced significantly with the emergence of multimodal methodologies that analyze facial expressions, vocal tones, and speech context to improve accuracy. Multimodal approaches address limitations associated with single-modal systems. Recent developments in the field of emotion recognition have opened new possibilities for building efficient and high-performance emotion detection systems. However, significant challenges remain in implementing these systems in real-time applications, achieving robust generalization across datasets and environments, and addressing different languages, such as Arabic and the Saudi Arabian dialect. Despite recent advances in emotion recognition, there are significant challenges in achieving reliable results. Emotions are complex and cannot always be accurately captured by a single feature, such as facial expressions or vocal tones. Relying on one modality can lead to misinterpretation, especially when facial expressions are unclear or vocal tones are affected by factors such as stress or the environment, so a multimodal approach is needed to combine features and improve accuracy.

Recent research has introduced a universal speech emotion recognition framework [26] called emotion2vec. Although promising, these frameworks have shown limited performance when evaluated on Saudi data. However, the field of facial emotion recognition is widespread. Many methods have been introduced. However, these approaches often ignore the role of vocal tones, leading to misinterpretations when facial expressions are unclear. This highlights the need for a multimodal approach that combines facial and vocal features for better accuracy. Many research studies have introduced a multimodal approach [24] [37]. Although previous studies have shown promising results with multimodal methods, they are often trained on invariant datasets, limiting their performance in different real-world environments. The objective of this study is to develop an audiovisual emotion recognition system that uses transfer learning for both image and audio frameworks designed for emotion recognition tasks [26] [8]. By combining facial and acoustic characteristics, the study aims to surpass state-of-the-art [25] on the benchmark RAVDESS [23] dataset and model performance on our collected data as mentioned in sectionIV. The audio arm is also trained on a Saudi Dialect dataset [1]. In this paper, we present our method to build a multimodal system and show the performance of the model in different environments [1].

II. Related Work

Multimodal emotion recognition has gained significant attention due to its ability to combine visual, audio, and textual data to achieve a more accurate understanding of human emotions. Recent advances in deep learning have enabled the integration of multiple modalities, opening the door to more reliable and adaptable emotion recognition systems. While challenges like combining different types of data remain, these advancements have the potential to change the way we understand and recognize emotions in everyday situations.

A. Speech Emotion Recognition (SER)

There are many attempts to only use acoustic features, this field is called speech emotion recognition (SER). SER has its roots in the late 1990s and early 2000s [38].

Many frameworks [11], [29], [30] use fine-tuned models that were created for speech tasks, such as Wav2vec [6], [34], WavLM [9], and Data2vec [4], [5].

One of the uses of WavLM [9] was in recent research [14], which explores leveraging content and acoustic representations for more efficient speech-emotion recognition. By combining two data types, the CARE [14] framework achieved a 65. 63%

[1] Source code is available on https://github.com/NaifMersal/MFER

average weighted f1 score. This score is the average computed on eight datasets and compared with six models/frameworks, where it achieved the highest score. It uses WavLM [9] and PASE+ [33] for the acoustic arm, and Whisper [32] and RoBERTa [22] models for the semantic arm. The eight datasets are spoken in English, limiting their flexibility and generalization to other languages such as Arabic. In response to language challenge, a recent study [26] introduced a universal expression representation model.

This model, Emotion2Vec, was pre-trained in 10 languages (e.g., Urdu, Persian, Greek), demonstrating consistent improvements across different languages in speech emotion recognition datasets. It generalizes well with SER tasks, such as song emotion recognition, emotion prediction in conversation, and sentiment analysis. Pre-trained in open source unlabeled emotion data through self-supervised online distillation, it combines utterance-level loss and frame-level loss, using the feature extractors of Data2vec 2.0 [4] and similar transformer and CNN architectures. Tested on 13 datasets and compared with four other models, Emotion2Vec outperformed all, including on the English language and others. However, it lacks training in Arabic data due to the absence of Arabic speech emotional datasets. Another model, VESPER [11], specific to the SER task, outperformed WavLM [9]. VESPER achieved an unweighted 78% average and a weighted 77% average.

Various architectures and methods have been explored in the SER field. One recent approach [13] proposes a Double Multi-Head Attention Multimodal System to address the limitations of conventional attention models. It removed CNN components and augmented the data, gaining a 10.1% validation Macro F1-score improvement on the MSP-Podcast dataset, although overall performance was unsatisfactory. Another approach aimed at improving classification accuracy [3] used the features of MFCC and a hybrid neural network combining CNN and ConvLSTM. Tested on the RAVDESS dataset with data augmentation, it achieved 91% accuracy after reducing the classes to six.

Furthermore, experiments for Arabic SER show promising results. One study [29] explored a deep learning-constructed emotion recognition model for Arabic speech dialogues, using Wav2Vec2.0 and HuBERT [18] representations on the BAVED [2] dataset. The model achieved 89% accuracy with pre-trained Wav2Vec2.0 [6]. Another study [1] used spoken data from the Arabic-Saudi dialect, applying SVM, MLP, and k-NN classifiers. The SVM achieved an accuracy of 77. 14%. Despite challenges in Arabic SER due to limited data, these efforts highlight significant advancements.

B. *Visual Emotion Recognition*

Visual emotion recognition takes advantage of facial and audio features to capture and analyze emotional expressions. Facial features, in particular, add significant dimensions to the recognition process. One paper [36] introduced the Spatiotemporal Convolutional Neural Network (ST-CNN) model for continuous emotion recognition using spatial and temporal aspects of facial and audio features. Using the SEWA-DB dataset, the model achieved 57% Arousal.

Another attempt [42] presented an End-to-End Visual-Audio Attention Network (VAANet) that focused on both audio and visual details. It used real-world video data from the VideoEmotion-8 dataset, achieving accuracy of 47. 5%. Two years later, a fully end-to-end system [39] balanced inference speed and recognition performance for emotion analysis in videos, achieving 79% accuracy on the IEMOCAP dataset.

A further study [43] introduced a Cross-Attention neural network with hybrid feature weighting, which improved the accuracy of emotion recognition on large video datasets, achieving 82% accuracy. However, this method relied on large-scale video data, which may limit its ability to cover all variations in emotions. Recently, cross-attention transformer fusion models, such as AVT-CA [37], combined audio and video data to achieve 83.2% accuracy. Nevertheless, high computational costs restrict its applicability in real-time.

The Vision Transformer (ViT) model [8] was also explored for visual emotion recognition tasks and compared with traditional models like ResNet-18. While challenges such as the size of the dataset and the distribution of the classes affected the results, ViT demonstrated potential, achieving 53% accuracy on datasets like FER-2013, AffectNet, and CK+48.

The article on Multimodal Emotion Recognition Using Aural Transformers and Action Units [24] was published in 2021, reaching a high accuracy on the RAVDESS dataset benchmark using a multimodal approach. For speech or voice emotion recognition (SER), a pre-trained transformer model, xlsr-Wav2Vec2.0 [12], is used, demonstrating significant progress in conventional speech-to-text conversion. For facial emotion recognition (FER), the system uses Action Units (AUs), which encode specific facial muscle movements associated with emotions. These are extracted using the OpenFace toolkit. Combining predictions through multinomial logistic regression resulted in an overall accuracy of 86.70%, significantly outperforming single models.

A mid-2024 study [16] explored single-modality and joint fusion approaches to emotion recognition on the RAVDESS dataset. It provided a comprehensive survey of methods across visual, speech, and audio-visual modalities. The findings highlighted that facial emotion recognition achieved better performance than speech-based methods, contrary to earlier studies. Furthermore, combining the modalities demonstrated the highest accuracy of 89%, underscoring the potential of joint fusion strategies.

A recent study [20], published in October 2024, proposed an audiovisual emotion recognition model leveraging bi-layer LSTM and multi-head attention for improved fusion of audio and visual features. Using the RAVDESS dataset, the model processed MFCCs for audio and facial features through convolutional layers, achieving an accuracy of 82.42% with 5-fold cross-validation.

Another notable contribution to visual emotion recognition [25], which introduces a framework that takes advantage of

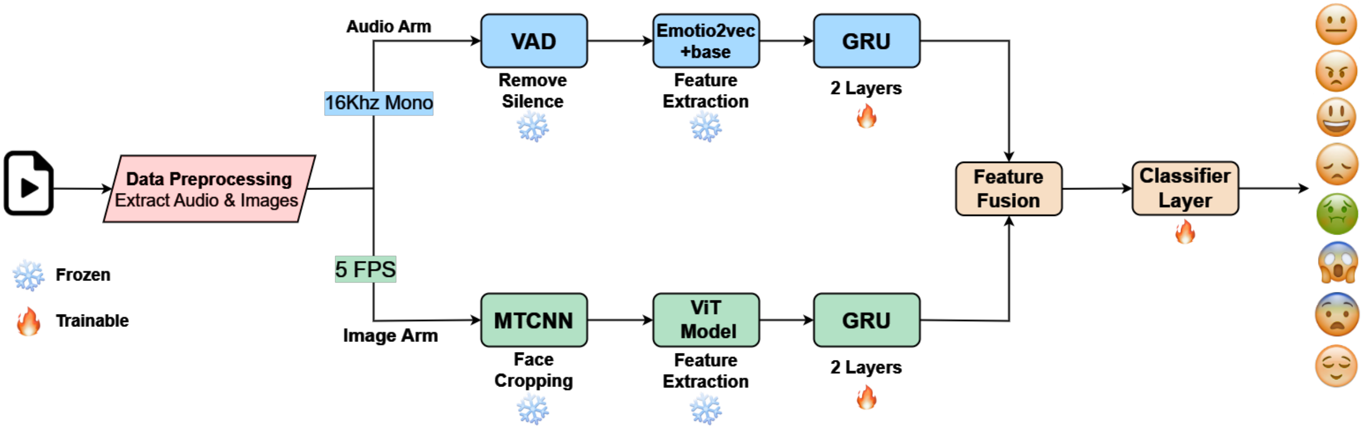


Fig. 1: Overview of the proposed multimodal architecture.

ResNet-50 [17] as the backbone architecture for feature extraction from both audio and visual modalities. The framework employs Hirschfeld-Gebelein-Rényi (HGR) maximal correlation loss for effective multimodal feature fusion, enhancing the stability and accuracy of emotion recognition. This approach achieved remarkable results on benchmark datasets such as RAVDESS [23], eNTERFACE'05 [27], and BAUM-1s [40].

C. State Of The Art (SOA)

Learning Better Representations for Audio-Visual Emotion Recognition with Common Information [25] article is a prominent contribution to this field. This paper introduces a novel framework designed to overcome key challenges in emotion recognition, such as efficient extraction of feature representations from multiple modalities and the use of common information.

The paper proposes a deep learning approach that integrates audio and visual modalities using correlation analysis. The methodology incorporates audio and visual networks, both of which utilize ResNet-50 [17] as their backbone architecture to extract feature representations. These features are then fused through a network trained with a joint loss function, combining correlation loss based on Hirschfeld-Gebelein-Rényi's (HGR) maximal correlation and classification loss. This innovative use of HGR maximal correlation allows the system to extract and utilize common emotional information across modalities, improving both feature stability and recognition accuracy. Furthermore, the framework is generalized to a semi-supervised learning scenario, enabling it to perform effectively even with limited labeled data.

We chose this study as the state-of-the-art because it introduces a novel multimodal approach that effectively blends audio and visual features, achieving an impressive accuracy of 97. 57% in the RAVDESS dataset. Recognizing that the split distribution can significantly affect results as demonstrated by our 98.6% performance on one fold, we then evaluated our approach using cross-validation.

The structure: The full loss function of the framework is a linear combination of correlation loss and classification loss. The fusion network has several fully connected layers. The correlation loss is used to extract common information between different modalities. Additionally, the classification loss is used to capture discriminative information from each modality for emotion prediction.

During the training process, emotion labels are used twice, once to calculate classification loss and the other as the third modality to compute correlation loss with audio and visual modalities. In this way, the label information can be fully used to improve the discrimination ability of the feature representations. In the testing process, audio and visual data are used to predict the corresponding emotion labels.

III. Methodology

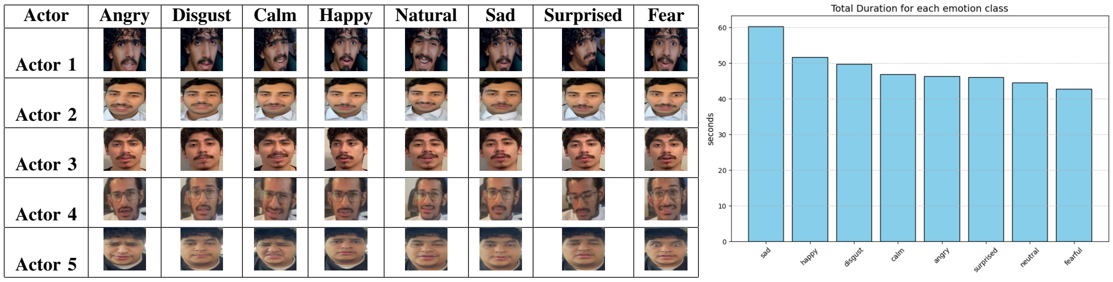
In our paper, we focus on implementing multimodal emotion detection. This system integrates two primary data: Audio and image data as in Figure 1. The goal is to make the system widely applicable in various sectors while improving the decision and reliability of emotion recognition.

A. Multimodal pipeline

Our multimodal consists of two primary arms; we merged the image and audio arms to predict the final emotion taking advantage of the facial and acoustic features. A video file is inputted into the model. Firstly, the video is processed by extracting the frames and audio waves. The wave-extracted will be sent to the audio arm, where they are processed as described in the audio arm section. The frames are sent to the image arm and processed as described in the image arm section. The output from each arm is processed through a fusion layer that concatenates the features extracted from both arms. The features are concatenated in a fusion layer, then passed through a fully connected layer, which performs the final classification task to determine the predicted emotion.

B. Audio Arm

First, the audio waves are normalized; after that, the audio enters Silero VAD to remove any silence by detecting speech and keeping only the parts of the audio clips that



(a) Actors' frame examples per emotion class.

(b) Total duration for each emotion class.

contain speech [35], then Emotion2vec [26] for extraction. The Emotion2vec plus base version is selected as a feature extractor because it is smaller than the large version. The features of Emotion2vec come as a vector with a dimension of 768. Approximately every 45ms, a vector is extracted. Then, it goes into two layers of GRU (768 → 384 dimensions) to learn the relationships in the audio over time. The GRU layers process the sequence of features, picking up on pitch and tone changes and then outputting them into a linear layer.

C. Image Arm

We start by extracting five frames per second from each video to capture key moments. These frames are processed using MTCNN [41], which detects and crops faces, resizes the images to 224x224 pixels, normalizes the pixel values and ensures that all images are in grayscale format. The cropped faces are then passed into a Vision Transformer (ViT) [19], which is modified to output feature vectors instead of predefined labels. These feature vectors represent each frame and are then entered into the GRU (768 → 384 dimensions) network, which looks at the sequence of frames to understand how emotions change over time. By combining MTCNN for precise face detection, ViT for detailed feature extraction, and GRU for tracking patterns across frames, we can accurately determine the overall emotion of the video.

IV. Data Collection and Techniques

The development of multimodality highlights the need to use a variety of datasets. The dataset consists of two primary components: an audio dataset for speech tasks and an image dataset for visual tasks.

A. Audio Data

We combined different datasets, aiming to have a variety of different tones, especially in the Arabic language and similar languages in terms of tones. We started by using the Saudi Dialect Corpus [1], which is made up of 175 records containing male and female actors, with 113 segments for males and 62 for females. The total duration of the dataset was about 11 minutes. Additionally, we used the URDU dataset [21], which contains emotional utterances of Urdu speech gathered from Urdu talk shows and 400 utterances of four basic emotions: Angry, Happy, Neutral, and Emotion. There are 38 speakers (27 male and 11 female). Targeting a similar tone to the Arabic tone, we used ShEmo [28], which covers speech samples of 87 Persian speakers for five basic emotions: anger, fear, happiness, sadness, and surprise. The database includes 3000 semi-natural utterances, equivalent to 3 hours and 25 minutes of speech data extracted from online radio plays. Eventually, for the sake of generalization, we used the CREMA-D dataset [7], which comes in the English language. It has 7,442 original clips from 91 actors.

TABLE I: Dataset Distribution Across Classes and Sources

Emotion	CREMA-D	SDC [1]	ShEMO	URDU	Total
angry	1271	69	1059	100	2629
fear	1271	0	38	0	1309
happy	1271	31	201	100	1715
neutral	1087	37	1028	100	2382
sad	1271	38	449	100	1985
Total	6171	175	2775	400	10020

B. Image data

On the other hand, the image model [19] we used was trained on the FER-2013 dataset [15], The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. It has seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples.

C. Audio-Visual Data

RAVDESS dataset [23] is ideal for visual audio tasks. It contains 24 professional actors (12 female, 12 male), vocalizing two matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprised, and disgusting expressions, and the song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

Additionally, we have created our own dataset, this dataset contains videos that include audio and visual data. The dataset is organized into 8 emotion classes, each class includes 3 predefined sentences, the sentences vary according to the class, performed by 5 actors, each actor reads the three sentences for each emotion class, and each video clip ranges from 2-4 seconds, the total clips in the dataset are approximately 45-60 for each class, this dataset provides an excellent representation of emotions in the context of the Arabic language. The statistics details are mentioned in Figure 2b

V. Training Strategy

Multimodal learning challenges the integration of various sensory inputs. We used a progressive training strategy for multimodal feature learning through a three-stage process. By implementing stratified cross-validation, we optimize model performance across visual and audio modalities.

The methodology sequentially trains the model using visual-only, audio-only, and combined modality stages, with defined hyperparameters. This approach allows the model to explore individual features first and then integrated features, addressing multimodal information processing.

A. Hyperparameters and Details

1) **Visual-Only Stage:**
   - Objective: Learn visual features independently
   - Used Data: RAVDESS [23]
   - Training duration: 30 epochs
   - Learning rate: 1×10^{-3} with exponential decay
   - Batch size: 16
   - Regularization: L2 weight decay (5×10^{-4})
2) **Audio-Only Stage:**
   - **First Stage:**
     - Objective: Learn audio features independently
     - Used Data: CREAMA-D [7], SheEMO [28], SDC [1], URDU [21]
     - Training duration: 30 epochs
     - Learning rate: 1×10^{-3} with on plateau scheduling
     - Batch size: 16
     - Number of classes: 4
     - Optimization: Adam optimizer
   - **Second Stage:**
     - Objective: Transfer learning and expanded classification
     - Base model: GRU from the first stage
     - Dataset: RAVDESS [23]
     - Number of classes: 8
     - Training approach:
       - Replace 4-class classifier with new 8-class classifier
       - Fine-tune on RAVDESS dataset
     - Learning rate: 1×10^{-3} for classifier
     - Learning rate: Reduced to 1×10^{-4} for GRU
     - Training duration: 25 epochs
     - Batch size: 16
3) **Combined Modality Stage:**
   - Objective: Fusion and joint representation learning
   - Used Data: RAVDESS [23]
   - Training duration: 8 epochs
   - Learning rate: 5×10^{-4} (reduced for fine-tuning)
   - Batch size: 16

VI. Results and Discussion

This section presents the performance outcomes for each arm, including the audio, video, and Multimodal models. All experiments were conducted as described in Section V.

A. Audio Model Performance with 4 classes

For the Audio Model, we evaluated its performance on various datasets, as shown in II:

TABLE II: Emotion Recognition Accuracy Across Datasets and Classes

Dataset	Emotion Class	Accuracy (%)
CREMA-D	**Angry**	**97.01** ± 0.47
	Happy	**94.02** ± 1.35
	Neutral	**96.41** ± 1.78
	Sad	82.91 ± 2.14
ShEMO	Angry	94.03 ± 2.00
	Happy	84.50 ± 4.00
	Neutral	91.41 ± 1.87
	Sad	75.28 ± 4.26
Saudi Dialect	Angry	73.85 ± 6.15
	Happy	70.00 ± 12.47
	Neutral	74.29 ± 10.69
	Sad	48.57 ± 14.57
URDU	Angry	97.00 ± 4.00
	Happy	90.00 ± 7.07
	Neutral	94.00 ± 4.90
	Sad	**91.00** ± 6.63

TABLE III: Comparison between MFER Audio Model and SVM on Saudi Dialect

Model	Results (%)
MFER Audio Arm	**67.88**
SVM [1]	77.0

The emotion detection system demonstrated strong performance across multiple datasets, with the URDU dataset showing the highest accuracy, closely followed by CREMA-D and ShEMO. This indicates that the model performs well on datasets with a balanced or clear emotional representation. However, the system struggled with the Saudi Dialect Corpus, likely due to variability in data quality, linguistic nuances, or cultural differences that may have influenced emotional expression. Among emotions, the model achieved the highest accuracy for the angry class, reflecting its ability to identify more distinct emotional cues in this category. Similarly, emotions such as happy and neutral were also recognized with high accuracy. However, the model faced challenges in detecting sadness, which can arise from subtle variations in speech patterns and tone, making it harder to distinguish it from other emotions. The combination analysis reveals nuanced insights. The best-performing combination was CREMA-D-angry, which demonstrates the system's strength in recognizing anger within CREMA-D's well-structured dataset. In contrast, the worst-performing combination was the Saudi Dialect Corpus, which highlights the difficulty in identifying sadness in the Saudi Dialect Corpus dataset. Even more, the SVM model in Table III achieved a higher accuracy of 77.0%, and our audio arm reached 67.88%. This suggests that the number of trained parameters is large compared to the amount of data in the dataset.

B. Multimodal Performance with 8 classes

Our approach performed exceptionally well on the RAVDESS dataset, with both the image and audio arms achieving comparable accuracy. By combining these two modalities, we achieved 96. 94%, outperforming multiple studies, as shown in Table VI. After cross-validating, we observed that while our results did not exceed the state-of-the-art [25] results, it is important to consider the differences in methodology. Unlike our approach, the state-of-the-art study which reported 97.57% accuracy, did not use cross-validation and variations in the data distribution could have influenced their outcomes. Interestingly, in one fold of our evaluation, we achieved an accuracy as high as 98.6%.

In contrast, when evaluating the generalizability of the model in our dataset (Table V), the multimodal resulted in unsatisfactory results. This performance degradation can be attributed to several factors, including inherent differences in dataset characteristics such as recording environments and cultural expression variations. Furthermore, the limited actor diversity in the RAVDESS dataset likely led to model overfitting, where the model memorized specific audio and facial features along with their changes over time. Notably, the audio model demonstrated marginally better performance compared to the visual and combined modalities, which can be partially explained by the prior training of the audio GRU on variant datasets, while the newly introduced classifier had insufficient opportunity to adjust weights properly.

TABLE IV: Performance Summary Across Modalities on RAVDESS Dataset

Modality	**Metric**	**Mean ± Std**	**Min**	**Max**
Visual	Accuracy	88.70 ± 1.82	86.38	91.16
Audio	Accuracy	89.20 ± 0.55	88.54	90.20
Combined	Accuracy	96.94 ± 1.27	94.79	98.61

TABLE V: Performance Summary Across Modalities on Our Dataset

Modality	**Metric**	**Mean ± Std**	**Min**	**Max**
Visual	Accuracy	18.8 ± 2.71	15.0	22.5
Audio	Accuracy	34.16 ± 7.16	20.83	41.66
Combined	Accuracy	26.14 ± 1.60	24.16	28.33

TABLE VI: Comparison of Multimodal Results on RAVDESS Dataset

Model	**Results (%)**
MFER Multimodal (Ours)	96.94
Aural Transformers and Action Units [24]	86.70
Joint Fusion [16]	89.00
bi-layer LSTM with Multi-Head Attention [20]	82.42
Common Information Framework [25]	**97.57**

VII. Conclusion

In this study, we set out to achieve several goals, and while we made progress, there were some challenges along the way. Testing our model on the benchmark RAVDESS dataset showed promising results, with performance comparable to that of the state-of-the-art and even exceeding it in certain folds. This demonstrates the strength of our approach and the power of transfer learning for emotion recognition. However, when applied to our own collected data, the results underscored the importance of having high-quality and diverse datasets to ensure consistent performance.

On the audio side, we managed to get good results for four emotion classes, but when it came to the Saudi dialect dataset, the results weren't as strong as we had hoped. The main issue was the limited amount of data, which also lacked variety, making it hard for our audio model to perform well.

One of the trickiest parts of this work was dealing with sad emotions, which often confused the model because they're so unclear and can be easily mistaken for neutral emotions. Our goal was to achieve high accuracy on a benchmark dataset while also using a local dataset for training. Although not all results met our expectations, this project opened the door to exciting possibilities for future work. The next steps could include using different audiovisual datasets to train the model [10], [27], [31], [40], refining our models to work better with limited data, and focusing on local data. These improvements could help create emotion recognition systems that are more adaptable and reliable for a different range of real-world applications.

References

[1] Reem Hamed Aljuhani, Areej Alshutayri, and Shahd Alahdal. Arabic speech emotion recognition from saudi dialect corpus. *IEEE Access*, 9:127081–127085, 2021.

[2] A. Aouf. Basic arabic vocal emotions dataset (baved) - github, September 2019.

[3] Youakim Badr, Partha Mukherjee, and Sindhu Madhuri Thumati. Speech emotion recognition using mfcc and hybrid neural networks. In *IJCCI*, pages 366–373, 2021.

[4] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language, 2023.

[5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. *CoRR*, abs/2202.03555, 2022.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[7] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[8] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. Vitfer: facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4):80, 2022.

[9] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[10] Sheng-Yeh Chen, Chao-Chun Hsu, Chia-Chun Kuo, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018.

[11] Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*, 2024.

[12] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020.

[13] Federico Costa, Miquel India, and Javier Hernando. Double multi-head attention multimodal system for odyssey 2024 speech emotion recognition challenge. *arXiv preprint arXiv:2406.10598*, 2024.

[14] Soumya Dutta and Sriram Ganapathy. Leveraging content and acoustic representations for efficient speech emotion recognition. *arXiv preprint arXiv:2409.05566*, 2024.

[15] Ian J Goodfellow, Dumitru Erhan, Pierre Laurent Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2013.

[16] Syrine Haddad, Olfa Daassi, and Safya Belghith. Single modality and joint fusion for emotion recognition on ravdess dataset. *SN Computer Science*, 5(6):669–, 2024.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[18] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[19] Dmytro Iakubovskyi. Facial emotions image detection, 2024. Copyright 2024 Dmytro Iakubovskyi dima806@gmail.com.

[20] Zeyu Jin and Wenjiao Zai. Audiovisual emotion recognition based on bi-layer lstm and multi-head attention mechanism on ravdess dataset. *The Journal of Supercomputing*, 81(1):31–, 2024.

[21] Siddique Latif, Adnan Qayyum, Muhammad Usman, and Junaid Qadir. Cross lingual speech emotion recognition: Urdu vs. western languages. In *2018 International conference on frontiers of information technology (FIT)*, pages 88–93. IEEE, 2018.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[23] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), April 2018.

[24] Cristina Luna-Jiménez, Ricardo Kleinlein, David Griol, Zoraida Callejas, Juan M. Montero, and Fernando Fernández-Martínez. A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. *Applied Sciences*, 12(1), 2022.

[25] Fei Ma, Wei Zhang, Yang Li, Shao-Lun Huang, and Lin Zhang. Learning better representations for audio-visual emotion recognition with common information. *Applied Sciences*, 10(20), 2020.

[26] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.

[27] O. Martin et al. Multimodal caricatural mirror. In *Proc. eNTERFACE 2005*, Mons, Belgium, 2005. Available on the eNTERFACE '05 website.

[28] Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami. Shemo: a large-scale validated database for persian speech emotion detection. *Language Resources and Evaluation*, 53:1–16, 2019.

[29] Omar Mohamed and Salah A Aly. Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset. *arXiv preprint arXiv:2110.04425*, 2021.

[30] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *CoRR*, abs/2104.03502, 2021.

[31] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Rada Mihalcea, and Erik Cambria. Meld: A multimodal multi-party dataset for emotion recognition in conversation. 2018.

[32] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[33] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition, 2020.

[34] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862, 2019.

[35] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

[36] Thomas Teixeira, Éric Granger, and Alessandro Lameiras Koerich. Continuous emotion recognition with spatiotemporal convolutional neural networks. *Applied Sciences*, 11(24):11738, 2021.

[37] Shravan Venkatraman, Vigya Sharma, Santhosh Malarvannan, et al. Multimodal emotion recognition using audio-video transformer fusion with cross attention. *arXiv preprint arXiv:2407.18552*, 2024.

[38] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814, 2021.

[39] Qinglan Wei, Xuling Huang, and Yuan Zhang. Fv2es: A fully end2end multimodal system for fast yet effective video emotion recognition inference. *IEEE Transactions on Broadcasting*, 69(1):10–20, 2022.

[40] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. BAUM-1. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5GK6C.

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[42] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 303–311, 2020.

[43] Siwei Zhou, Xuemei Wu, Fan Jiang, Qionghao Huang, and Changqin Huang. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. *International Journal of Environmental Research and Public Health*, 20(2):1400, 2023.