



**AIN SHAMS UNIVERSITY**

**FACULTY OF ENGINEERING**


**Computer Engineering and Software Systems Program - International Credit Hours Engineering Programs (I-CHEP)**

# Multi-Modal Emotion Recognition

---

**Under the Supervision of:**

*Dr. Mohamed Rehan*

**Student's names:**


| | |
|---|---|
| *Abdel Rahman Mohamed Salah* | *19P9131* |
| *Mahmoud Mohamed Seddik Hassan ElNashar* | *19P3374* |
| *Mohamed Hatem Zakaria Elafifi* | *19P7582* |
| *Omar Ashraf Mabrouk Abdelelwahab* | *19P8102* |
| *Omar Mohamed Ibrahim Alsayed* | *19P7813* |
| *Zakaria Sobhy Abd-Elsalam Soliman Madkour* | *19P2676* |

**Faculty of engineering Ain Shams University**
**International Credit Hour Program**
**Computer Engineering and Software Systems Program**

# Graduation Project Proposal Form

| | |
|---|---|
| **Project Serial** | # To be completed by the Course Coordinator |
| **Project Title** | **Multimodal Emotion recognition system** |
| **Supervisor(s)** | Dr. Mahmoud Khalil |
| **Teacher Assistant (if any)** | N/A |
| **Sponsoring Company (if any)** | N/A |
| **Number of Students** | 6 |

**Names/IDs of Students**

| | |
|---|---|
| 1. Abdel Rahman Mohamed Salah | 19P9131 |
| 2. Mahmoud Mohamed Seddik Hassan ElNashar | 19P3374 |
| 3. Mohamed Hatem Zakaria Elafifi | 19P7582 |
| 4. Omar Ashraf Mabrouk Abdelelwahab | 19P8102 |
| 5. Omar Mohamed Ibrahim Alsayed | 19P7813 |
| 6. Zakaria Sobhy Abd-Elsalam Soliman Madkour | 19P2676 |

**Project Description**

The problem under study in this project **is the impact of combining the results of both visual/speech emotion detection results** to achieve better overall result from the model. We applied it in the evaluation of **employees' performance in customer service interactions**. The project aims to utilize visual/speech emotion detection techniques to analyze the emotions displayed by the representative **throughout a video call or face to face interaction**. The project will encompass emotions expressed by individuals. The findings from this analysis can **provide valuable insights into areas for improvement** in customer service, **contributing to the enhancement of the overall customer experience**.

**Proposed Datasets**

For visual ED: CK+, FER+, RAF, GoEmotions (Examples not limited to them).
For Speech ED: Crema, Savee, Tess, Ravee (Examples not limited to them).

**Project Objectives**

- Develop and implement visual and speech emotion detection models.
- Utilize different datasets with extensive visual and speech data, encompassing multiple emotions expressed by numerous individuals.
- Integrating multiple datasets from different sources with the suitable preprocessing to achieve generalization and increased volume.
- Implementing deep learning models for visual emotion detection and if needed for the speech part.

- By employing model fusion techniques, leverage the strengths of both visual emotion detection and speech emotion detection by combining their results.
- Explore the correlation between specific emotions and changes in representative interactions levels.
- Evaluate the overall behavior of the representative during the interaction.
- Provide actionable insights and recommendations for improving customer service based on the analysis of emotions and behavior.

## Required Prior Skills

- Familiarity with **deep learning** frameworks like TensorFlow or PyTorch.
- Proficiency in **computer vision** techniques is essential to process and analyze visual data. This includes image preprocessing, feature extraction, object detection, and image classification using CNN architectures like ResNet, VGG, or Inception.
- Skills in **preprocessing data from different sources**, normalizing features, handling missing values, and standardizing data formats are necessary. Additionally, integrating data from visual and speech modalities requires expertise in feature alignment and fusion techniques.
- Knowledge of various **model fusion** techniques is crucial to effectively combine the outputs of visual and speech emotion detection models.
- Proficiency in Python and related **libraries** (NumPy, Pandas, scikit-learn) is essential for implementing machine learning algorithms, data manipulation, and model development.
- Understanding **evaluation metrics** specific to emotion detection, such as accuracy, precision, recall, F1-score, or confusion matrices, is important for assessing the performance of combined models.
- Familiarize ourselves with **existing research**, methodologies, and algorithms related to visual and speech emotion detection.
- Familiarize ourselves with **relevant datasets** available for emotion detection in customer service interactions.
- Develop an understanding of customer service dynamics and factors that influence customer satisfaction.
- Study and review research papers or case studies on customer service improvement strategies and techniques.

## Deliverables with estimated time plan (Semester 2)

- **-Preprocess** and clean the acquired speech emotion datasets.
  **-Explore** various audio processing methods and feature extraction techniques for speech emotion detection.
  (Week 1-2)

- **-Implement** and train a speech emotion detection model using one of the selected datasets.
  **-Evaluate** and fine-tune the performance of the speech emotion detection model.
  (Week 3-4)

- **-Integrate** the speech emotion detection model with the visual emotion detection model developed in the first semester.
  (Week 5-7)

- **-Develop** a user interface for the application to display combined results.
  (Week 8-10)

- **-Implement** additional features such as data visualization, performance metrics, and user management in the application.
  **-Test and validate** the application's functionality and performance.
  (Week 11-14)

- **-Conduct** thorough testing and debugging of the combined application.
  **-Prepare** comprehensive documentation, including user guides and technical manuals.
  **-Finalize** the project, conduct a final evaluation, and present the complete work.
  (Week 15-16)

## Other potential applications for our idea:

- Chrome Extension for Websites Ranking/Categorizing
- Driver Safety and Alertness
- Mental Health and Wellness
- Games emotional impact assessment
- Smart home, Mood-Based Music Recommendation

## ABSTRACT:

Emotion detection is a critical aspect of human-computer interaction and artificial intelligence applications. This bachelor project explores the combination of preexisting visual and audio models to enhance the accuracy of emotion detection systems. Rather than developing models from scratch, we utilize advanced visual and audio models and investigate their synergies for a more comprehensive understanding of human emotions.

The project examines the interaction between facial expressions, extracted through established computer vision algorithms, and audio features derived from signal processing methods. Notably, this study adopts a strategy of merging a preexisting visual model with an audio model, aiming to capitalize on the strengths of each modality and achieve improved overall accuracy.

Facial landmarks are extracted using cutting-edge computer vision techniques, and audio features such as pitch, intensity, and tempo are analyzed. Machine learning models, including deep neural networks, are then utilized to combine the outputs of the preexisting visual and audio models. The project assesses the impact of this multimodal fusion on the enhancement of emotion detection accuracy.

Results in theory should indicate a significant improvement in accuracy compared to individual modalities, highlighting the efficiency of combining preexisting models. The integrated system demonstrates robustness across diverse emotional expressions and speech patterns. The project underscores the pragmatic approach of combining established visual and audio models to achieve heightened accuracy without reinventing the wheel.

In conclusion, this project advances the field of emotion detection by presenting a pragmatic approach that combines preexisting visual and audio models. The findings emphasize the effectiveness of this integration, offering a pathway for more sophisticated and accurate applications in artificial intelligence and human-computer interaction.

# TABLE OF CONTENTS

Abstract: ..... i
List of Figures: ..... iv
List of Tables ..... iv
List of Abbreviations ..... iv
Chapter One: Introduction ..... 1
Motivation ..... 1
Problem Statement ..... 2
Outline ..... 3
Chapter Two: Literature Review ..... 4
Datasets and Benchmarks ..... 4
Popular Datasets for Multimodal Emotion Detection ..... 4
Benchmarks and Challenges ..... 5
Impact on Model Development ..... 6
Video Data Availability: ..... 7
Fusion techniques review ..... 9
Wav2vec2.0 (acoustic encoder) ..... 9
RoBERTa$_{BASE}$ (contextual encoder) ..... 10
MMER (Multimodal Multi-task Learning for speech emotion recognition) ..... 11
Efficient Face ..... 12
Chapter Three: NN Architecture Review ..... 13
Visual Model ..... 13
i. FaceNet: ..... 13
ii. EfficientFace: ..... 16
Audio Model ..... 19
i. Wav2Vec ..... 20
ii. RoBERTa$_{BASE}$ ..... 20
Chapter Four: Solution Architecture ..... 21
Project Timeline ..... 21
Visual Model (FaceNet and EfficientFace Integration) ..... 21
Audio Model (MMER with Wav2Vec2 and RoBERTaBASE) ..... 22
Fusion Implementation ..... 23

Method 1 ..... 24
Method 2 ..... 25
Method 3 ..... 26
Method 4 ..... 27
Method 5 ..... 28
Chapter Five: Software Requirements Specifications (SRS) ..... 29
Scope ..... 29
Functional Requirements ..... 30
Non-Functional Requirements ..... 31
Performance Requirements ..... 31
- Training Time: The training process for both the efficient face model and the Wave2Vec2 audio model should be completed within a reasonable time frame, considering computational resources. ..... 31
- Inference Time: The Emotion Detection Model should provide predictions for new input data in real-time or within acceptable response time limits. ..... 31
Reliability ..... 31
- Robustness: The Emotion Detection Model should handle variations in input data gracefully, providing consistent and reliable predictions across different scenarios. ..... 31
Scalability ..... 31
- Model Enhancement: The system should be designed to accommodate future enhancements, including the integration of additional modalities or improvements to existing models. ..... 31
- Dataset Variations: The Emotion Detection Model should demonstrate scalability in handling variations in input data, such as diverse facial expressions and emotional nuances in speech and song. ..... 31
Use Cases ..... 32
Chapter Six: Prototype Results ..... 34
Chapter Seven: Conclusions ..... 35
References ..... 36

## LIST OF FIGURES:

*Figure 1: RAVDESS dataset distribution* ..... 7
*Figure 2: FaceNet architecture. You can see the full details of the layer's sizes, The flow of the data and PARM, FLOPS in Table2* ..... 14
*Figure 3: Effiecientface abstract architecture showing its pipeline.* ..... 16
*Figure 4: Local-feature extractor used in EffecientFace components.* ..... 17
*Figure 5: Channel-spatial modulator compnents* ..... 18
*Figure 6: (a) Represents the normal residual block which has a wide -> narrow -> wide structure with the number of channels. (b) Inverted Residual Block follows a narrow -> wide -> narrow approach, hence the inversion.* ..... 18
*Figure 7: MMER architecture* ..... 19
*Figure 8: Cross Modal Encoder* ..... 19
*Figure 9: Wav2Vec2 ver2.0 pipeline.* ..... 20
Figure 10 Fusion 1 ..... 24
Figure 11 Fusion 2 (removed last layer of each model) ..... 25
Figure 12 Fusion 3 (Used Average Pool instead of FC in visual model) ..... 26
Figure 13 Fusion 4 (Using ELU and Dropout) ..... 27
Figure 14 Fusion 5 (Using ELU, Dropout, and Max Pooling) ..... 28

## LIST OF TABLES

*Table 1: Number of Annotated Images in Each Category* ..... 8
*Table 2: table show the structure of our Zeiler&Fergus based model with 1×1 convolutions. The input and output sizes are described in rows × cols × #filters. The kernel is specified as rows × cols, stride and the maxout pooling p=2* [4]. ..... 15
*Table 3: Performance of different fusion methods on RAVDESS DB compared to the original models.* ..... 34

## LIST OF ABBREVIATIONS

| | |
|---|---|
| *MMER* | (Multimodal multi-task learning approach for Speech Emotion Recognition). |
| *FER* | Facial emotion recognition |
| *IEMOCAP* | Interactive Emotional Dyadic Motion Capture |
| *RAVDESS* | Ryerson Audio-Visual Database of Emotional Speech and Song |
| *MELD* | Multimodal Emotion Lines Dataset |

# CHAPTER ONE: INTRODUCTION

## Motivation

The motivation behind this research is deeply embedded in the transformative landscape of artificial intelligence, which permeates diverse facets of human life, reshaping our interactions with technology. As AI systems progressively assimilate into daily routines, the recognition of the vital role that emotions play in human-computer interaction becomes increasingly apparent. Emotion detection stands as a critical linchpin in achieving a harmonious integration of AI into our lives, not only promising enhanced user experiences but also fostering the development of empathetic and contextually aware applications.

While existing emotion detection systems have made notable strides in both visual and audio-based recognition, inherent limitations persist. Unimodal approaches, whether reliant on facial expressions or audio cues, often struggle to encapsulate the intricate tapestry of human emotions, inherently multimodal in nature. It is imperative to note that while there are existing models attempting to combine both modalities, they frequently exhibit suboptimal accuracy, underscoring the need for a more effective and unified strategy.

Furthermore, with the continuous evolution of AI technologies, there is a growing imperative to bridge the gap between human emotions and machine understanding. Applications ranging from virtual assistants to gaming environments and mental health support systems could significantly benefit from emotion detection systems that not only exhibit heightened accuracy but also possess contextual awareness. The motivation driving this research is deeply rooted in the pursuit of a pragmatic solution that acknowledges and addresses the limitations of current models, paving the way for a new era of emotionally intelligent AI that transcends the shortcomings of existing unimodal approaches.

## Problem Statement

The predominant challenge in prevailing emotion detection methodologies stems from their reliance on single-modal approaches, exclusively emphasizing either facial expressions or speech patterns. These unimodal systems often prove insufficient in capturing the intricate and multifaceted nature of human emotions, as they disregard the potential synergies between visual and auditory cues. Consequently, the present models may misinterpret or overlook subtle emotional nuances, diminishing accuracy and reliability in real-world applications.

Moreover, unimodal approaches encounter difficulties in scenarios where one modality provides ambiguous or incomplete information. For instance, in noisy environments or situations where facial expressions are obscured, relying solely on visual cues may lead to inaccuracies. Similarly, exclusive dependence on audio signals may face challenges in contexts where speech is absent or lacks discernible emotional cues.

Importantly, it's worth noting that while some models attempt to combine both modalities, their accuracy remains suboptimal. In real-world applications, these integrated models may still struggle to provide reliable emotion detection results. The identified limitations underscore the imperative need for a multimodal approach—one that integrates preexisting visual and audio models to harness the distinct strengths of both modalities. By merging the capabilities of facial expression recognition with the nuances embedded in speech patterns, a more comprehensive and accurate emotion detection system can be realized. This research endeavors to address this critical gap by exploring the integration of preexisting visual and audio models, aiming to contribute to the advancement of emotion detection technologies and their practical applications.

## Outline

**Literature Review:**

- Comprehensive examination of existing emotion detection models, exploring unimodal and multimodal approaches.
- Identification of gaps in current research to establish the need for an integrated visual and audio model.

**Neural Network Architecture of utilized Preexisting Models**

- In-depth analysis of neural network architectures utilized in established visual and audio emotion detection models.
- Evaluation of the strengths and limitations of these architectures as a foundation for the proposed solution.

**Proposed Solution Architecture**

- Detailed exposition of the architecture proposed for integrating visual and audio models in emotion detection.
- Discussion of the methodology for combining facial landmarks from computer vision and audio features through signal processing.
- Integration of deep neural networks for multimodal fusion and enhanced accuracy.

**System Requirements Specification (SRS)**

- Documentation of functional and non-functional requirements essential for system development and deployment.
- Definition of the system's scope, specifying supported emotional expressions, environmental conditions, and user scenarios.

**Results**

- Evaluation of the integrated system's performance compared to individual visual and audio models.
- Discussion of findings, challenges encountered during implementation, and implications for future research and applications.

**Conclusion**

- Summation of key findings and contributions to the field.
- Reflection on the proposed solution's effectiveness in advancing emotion detection in AI systems.
- Recommendations for further research and applications, emphasizing the project's role in bridging gaps in current emotion detection methodologies.

# CHAPTER TWO: LITERATURE REVIEW

## Datasets and Benchmarks

Popular Datasets for Multimodal Emotion Detection

### 1- IEMOCAP (Interactive Emotional Dyadic Motion Capture):

**Modalities**: Speech, Facial Expressions.

**Description:**

- Consists of 151 recorded dialogues with 2 speakers per session, totalling 302 videos.
- Each segment is annotated for the presence of 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral).
- Additional annotations include valence, arousal, and dominance.
- Recorded across 5 sessions with 5 pairs of speakers.

**Structure:** Likely organized into folders representing sessions, each containing subfolders for speaker pairs.
Within each subfolder, video files (dialogues) would be present.
Associated annotation files or metadata providing details on emotions, valence, arousal, and dominance.

### 2- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):

**Modalities:** Speech, Facial Expressions.

**Description:**

- Contains 7356 files with a total size of 24.8 GB.
- Includes 24 professional actors (12 female, 12 male) vocalizing statements in a neutral North American accent.
- Expressions cover emotions such as calm, happy, sad, angry, fearful, surprise, and disgust for speech and song.

**Structure:** Organized into folders based on actor information and modality (speech or song).
Subfolders for each emotion within the speech and song modalities.
Each subfolder likely contains audio files corresponding to actor and emotion.

### 3- MELD (Multimodal Emotion Lines Dataset):

**Modalities:** Speech, Facial Expressions.

**Description:**

- Multimodal dataset with 13,118 utterances from 1,433 dialogues between 804 speakers.
- Annotated with emotion labels and sentiment scores.

**Structure:** Folders representing dialogues, each containing subfolders for speakers.
Subfolders may further categorize data based on emotions or sentiment.
Within each subfolder, audio files, transcripts (language), and possibly visual cues.

### 4- EmoReact:

**Modalities:** Speech, Facial Expressions.
**Description:**

- Contains 1102 audio-visual clips annotated for 17 different emotional states, including six basic emotions, neutral, valence, and nine complex emotions.

**Structure:** Likely organized into folders based on emotional categories.
Each folder contains audio-visual clips corresponding to different emotional states.
Metadata or annotation files providing details on the labeled emotions.

### 5- CMU-MOSEI (Multimodal Sentiment Intensity):

**Modalities:** Speech, Facial Expressions.
**Description:**

- Dataset with 93 hours of video and audio data, including 23,453 utterances.
- Annotated with sentiment and emotion intensity scores.

**Structure:** Organized into folders or sessions based on video and audio data.
Likely subfolders for individual utterances or speakers.
Annotation files providing sentiment and emotion intensity scores.

Benchmarks and Challenges

**1. Emotion Recognition Challenges:**

- RAVDESS has been employed in various emotion recognition challenges and competitions, serving as a benchmark for evaluating the performance of models across different emotion categories.
- Challenges often entail tasks such as classifying emotions in both speech and song performances, highlighting the versatility of the dataset in assessing model robustness.

**2. Contributions to Research:**

- RAVDESS has played a pivotal role in advancing emotion detection research, particularly in the realm of audio-visual emotion analysis.
- Researchers have utilized the dataset to explore the effectiveness of models in capturing emotions through both auditory and visual cues, contributing to the development of more sophisticated and accurate emotion recognition systems.

**3. Cross-Modal Studies:**

- The synchronized audio and visual data in RAVDESS facilitate cross-modal studies, encouraging researchers to develop models capable of integrating information from both modalities.
- Cross-modal studies are instrumental in understanding how emotional information is conveyed and perceived through multiple channels, leading to more comprehensive multimodal models.

Impact on Model Development

- Enhanced Model Robustness:

  The inclusion of both speech and singing performances, along with diverse emotional expressions, enhances the robustness of models trained on RAVDESS.

  Models developed on this dataset are better equipped to handle a wide range of emotional scenarios.

- Comparative Analysis:

  RAVDESS facilitates comparative analyses between unimodal (audio-only or visual-only) and multimodal models, allowing researchers to assess the benefits of combining information from multiple modalities.

- Real-world Applicability:

  The dataset's focus on naturalistic emotional expressions contributes to the development of models with real-world applicability, particularly in scenarios where emotions are expressed through speech or singing.

- Innovation in Model Architectures:

  Researchers working with RAVDESS are encouraged to explore innovative model architectures that can effectively fuse information from audio and visual modalities, pushing the boundaries of multimodal emotion detection.

## Video Data Availability:

Although all three datasets we came across while working on this project are big in size (in Gigabytes) they still contain few data. For example, IEMOCAP contains only 5 conversations, RAVDESS contains 24 actors only **Figure 1** . The area of emotion detection using video feed faces a huge hurdle due to a scarcity of video-based datasets, making it difficult to train viable algorithms. Despite the growing relevance of understanding and interpreting human emotions, there is a noticeable lack of extensive video datasets designed specifically for emotion identification tasks.

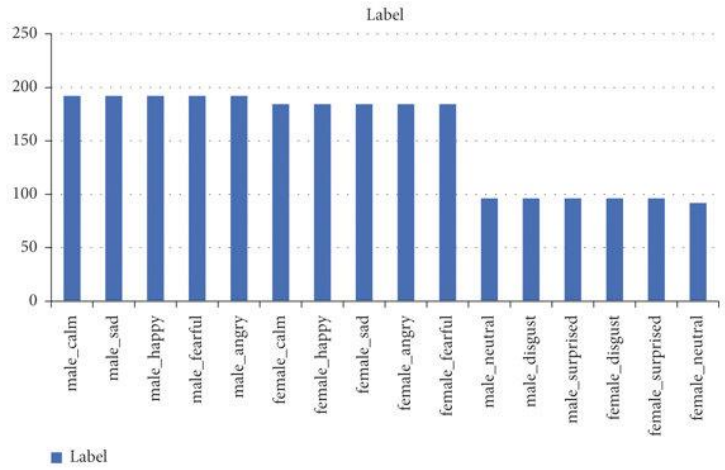


*Figure 1: RAVDESS dataset distribution*

Existing datasets that address this topic are frequently small, with only a few samples. The lack of varied and comprehensive video datasets impedes the creation and assessment of reliable emotion identification algorithms, stalling progress in this critical field of study. The lack of data not only limits the ability to create advanced algorithms, but it also impedes the investigation of subtle emotional responses across varied demographic groups and cultural situations. Addressing this gap in the availability of high-quality video datasets is critical for improving the capabilities of emotion identification algorithms and promoting their practical application in real-world settings.

To solve the difficulty of data scarcity for video-based emotion recognition (VER), we combined two powerful modalities: visual and auditory with some work arounds which are:

1. Using the capabilities of EfficientFace for visual emotion recognition (ER), an image-based approach, we were able to access substantially bigger picture datasets. EfficientFace was pretrained on AffectNet7. AffectNet contains more than 1M facial images collected from the Internet by querying three major search engines using 1250 emotion related keywords in six different languages [1]. To improve its temporal knowledge, we fine-tuned the model with the RAVDESS dataset, which, being a video database, naturally integrates temporal dynamics. This not only increased the amount of visual data available for training, but also allowed the model to detect subtle emotional expressions over time.

| Expression | Number |
|---|---|
| Neutral | 80,276 |
| Happy | 146,198 |
| Sad | 29,487 |
| Surprise | 16,288 |
| Fear | 8,191 |
| Disgust | 5,264 |
| Anger | 28,130 |
| Contempt | 5,135 |
| None | 35,322 |
| Uncertain | 13,163 |
| Non-Face | 88,895 |

*Table 1: Number of Annotated Images in Each Category*

2. In addition, Connectionist Temporal Classification (CTC) loss function. The use of CTC loss helps to alleviate data scarcity problems by allowing the model to learn from data without needing explicit matching with labelled emotion segments. Our methodology seeks to overcome data restrictions by utilising a multimodal approach and smart modifications [2]. In supervised learning contrastive loss cannot handle many samples belonging to the same class owing to labelling. When the number of positives is increased, additional functionalities become available equations (1),(2).

$$L_{out}^{sup} = \sum_{i \in I} L_{out,i}^{sup} = \sum_{i \in I} -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\frac{zi \cdot zp}{\tau}}}{\sum_{a \in A(i)} e^{\frac{zi \cdot za}{\tau}}} \quad \textbf{(1)}$$

$$L_{in}^{sup} = \sum_{i \in I} L_{in,i}^{sup} = \sum_{i \in I} -\log\{\sum_{i \in I} -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\frac{zi \cdot zp}{\tau}}}{\sum_{a \in A(i)} e^{\frac{zi \cdot za}{\tau}}}\} \quad \textbf{(2)}$$

## Fusion techniques review

In their papers [3] and [4], the writers described briefly the different methods of modalities fusion. The problem of multiple modality fusion could be seen as four smaller sub problems as follows: single-modal feature extraction, feature fusion, model classification, and result output. Depending on when fusion takes place between the stated sub stages we can classify fusion into three categories namely early fusion, late fusion, and intermediate fusion. Early fusion takes place when different modalities are fused by a series of dot products or concatenation of features.

The other extreme is to have multiple models one for each modality and fuse between the classification outputs of all the modalities models, which is known as late fusion. Intermediate fusion occurs when each modality runs a feature extraction network then the fusion happens between the extracted features before classification.

It has been widely believed that intermediate fusion is the best approach to address the problem of multiple modalities especially after the effectiveness of deep learning architectures such as CNNs and fusion models. Based on our reading in this topic we decided to apply both late, and intermediate fusion, since it was reported that late fusion in most of the cases yielded good results.

## Wav2vec2.0 (acoustic encoder)

Wav2vec2 has gained its unprecedented leadership in the field of processing raw audio data owing to self-supervised learning, and transformer technology. It is designed to address fundamental audio processing challenges such as feature extraction, contextual understanding, and representation learning. Aided with self-supervised learning wav2vec2 was able to be trained on much larger data corpus since it is able to be trained directly on raw audio waveforms without the need of explicit annotations.

1. Feature extraction

Wav2vec2 is feed with raw input audio signal (sequence of amplitudes over time) that passes through multi-layer convolutional neural network that extracts low-level features that capture the primitive acoustic speech properties.

2. Contrastive learning

The goal of contrastive learning isn’t classification, but, learning better representations of the acoustic signal in the latent space. Here is how it works: the model creates multiple representations in the latent space per audio segment. These different representations are either similar or dissimilar to the given segment. The goal of contrastive learning is to pull similar latent representations near together, and push

dissimilar pairs away from each other. This way the latent representation is robust meaningful representation of the audio segments.

3. Encoder

This part of the model is composed of the encoder part of a transformer which is feed with the low-level extracted features from the CNN and then uses these features to draw long-range dependencies between different segments within the same audio file.

4. Output layer

Based on the task this output layer will vary accordingly. If the task is automatic speech recognition (ASR) then this layer will be a decoder that outputs the highest probability sequence of textual words representing this acoustic file. In our case we will either use this layer as an emotion classifier where it is trained on one of the famous emotions datasets and then utilize this output emotion in the late fusion layer, or use the latent representation of the transformer encoder for the intermediate fusion layer.

## RoBERTa$_{BASE}$ (contextual encoder)

BERT stands for Bidirectional Encoder Representation from Transformers which is a natural language proceeding model. It is used to grasp the meaning of each word in a sentence. BERT utilizes bidirectional context understanding, and parallelized processing of a sequence thanks to the transformer architecture.

BERT analyses a phrase as a whole taking into account preceding and following words, in contrast to other unidirectional sequence models such as RNNs. This gives BERT the power to understand the meaning of the word in its context. Take for example the sentences “Teddy bears are on sale.”, and “Teddy Roosevelt was a great President.” the word Teddy changes its meaning depending on the context of the words following it. The bidirectional nature of BERT’s architecture is capable of capturing such difference in the representation of the word teddy depending on the sequence that follows.

BERT's parallelization is one of its main benefits over RNNs, its predecessor. This implies that although RNNs must process a sentence one word at a time, BERT may process a sentence as a whole at once. This results in great efficiency in its training time. Furthermore, because BERT is parallelized, it can capture long-range relationships between words more readily than RNNs, which are only able to consider a small amount of context at once.

## MMER (Multimodal Multi-task Learning for speech emotion recognition)

In their paper [5] Sreyan Ghosh, et al. tried to extract detailed multimodal emotional data from text and audio modalities by utilizing voice and the text transcripts that go along with it. In order to discover the natural harmonic alignment between voice and text as well as the syntactic and semantic information concealed in the text, it first resolves the problem of automatic speech recognition ASR by minimizing the CTC loss.

They state that most of the fusion model between the speech and text modalities were late fusion and they yeilded good results, yet this late fusion technique failed to capture cross-modal interactions. They also mention that although early fusion solves this proble that was posed by the late fusion method, they experience suppresion of modality-specific interactions. They suggest that a special type of intermediate fusion namely cross-modal attention mechanism where features from one modaliity could attend to features from the other modality.

The suggested approach to handle this task was as follows:

1. Features encoding: they used wav2vec2 trained on 960 hours of Librispeech as an acoustic encoder, and $\text{Roberta}_{\text{Base}}$ as the text encoder.

2. Multimodal interaction module:

This module contains three cross-modal encoders (CME) to capture the interactions between text, and speech representations. The **speech-aware word representation** CME takes wav2vec2 embeddings as queries and Roberta as keys and values. Similarly the **word-aware speech representation** CME takes Roberta embeddings as queries, and wav2vec2 embeddings as keys and values. **Acoustic gate** CME dynamically control the contribution of each speech frame embedding to get rid of the effect of random noise encoding.

Finally, after concatenation, sigmoid gate, and element-wise multiplication of the outputs of these three CME, this embedding is used to draw an emotion for the given text, and audio recording.

Results show that this approach is the state-of-the-art for performing emotion analysis on the IEMOCAP dataset with accuracy of 81.2%

## Efficient Face

EfficientFace is a facial expression recognition (FER) model that is meant to be parameter-efficient while maintaining accuracy and resilience in recognising facial emotions under a variety of situations, such as those found in real-world settings. To improve the resilience of the lightweight network, the authors include a local-feature extractor and a channel-spatial modulator. The inclusion of depth wise convolution in these components allows the network to record both local and global salient face information, increasing sensitivity to nuanced facial emotions.

In addition to architectural improvements, the study proposes a new training technique termed the Label Distribution Learning (LDL) method. This approach recognises the complexity of emotions, noting that they frequently present as combinations, mixes, or compounds of fundamental emotions. The LDL technique allows the model to learn not only individual emotion classes, but also their connections and combinations. This allows for a more nuanced and thorough grasp of the complicated nature of complex emotional responses. The proposed EfficientFace is rigorously tested on realistic occlusion and posture variation datasets to demonstrate its durability under demanding settings. The experimental findings reveal state-of-the-art performance on benchmark datasets such as RAF-DB, CAER-S, and AffectNet-7, with excellent accuracies of 88.36%, 85.87%, and 63.70%. Furthermore, the model gets a competitive performance on the AffectNet-8 dataset, with an accuracy of 59.89%.

To reduce computational overheads, the authors use the cutting-edge lightweight network ShuffleNet-V2 as a backbone network, which includes Conv1, Stage 2, Stage 3, Stage 4, and Conv5. The model incorporates a local-feature extractor and a channel-spatial modulator to address real-world difficulties like occlusion and posture variation.

The **local-feature extractor** is intended to effectively learn local face characteristics without relying on facial landmarks, hence solving the inefficiency of current approaches. Following Conv1, the global feature maps are divided into four patches, with each patch undergoing depthwise convolution to produce local face features. These local characteristics are subsequently concatenated, increasing the network's sensitivity to facial expressions associated with action units.

To reduce duplicate information in global face characteristics, a **channel-spatial modulator** is added after Stage 2. This modulator, modelled after the BAM architecture, emphasises critical global aspects. Channel and spatial heatmaps are generated, normalised, and merged to modify global characteristics. The final global-local features are created by mixing extracted local features with modulated global features, allowing the network to learn both global-salient and local face characteristics.

# CHAPTER THREE: NN ARCHITECTURE REVIEW

## Visual Model

Before we start talking about the architecture of the visual model lets first break it down into three main components: Facenet which is used for face detection, Efficientface which is the base of the visual emotion recognition model using images, and the final model which adds the capability to use videos or sequences of images.

### i. FaceNet:

Although we use it only for face detection FaceNet is a face recognition system that allows for fast face verification, identification, and grouping at scale(clustering). The model learns to map face pictures into a compact Euclidean space, where distances directly indicate the similarity of faces. It does this by training to optimise a 128-dimensional embedding using a new triplet-based loss [3] function. The network is trained using triplets of generally aligned matching and non-matching face patches obtained by online triplet mining. FaceNet's main benefit is its representational efficiency, which allows it to achieve cutting-edge face recognition performance while consuming just 128 bytes per face. Unlike prior efforts, FaceNet simply trains its output to be a compact embedding, eliminating the need for an intermediate bottleneck layer and resulting in a far more efficient representation per face.

**Triplet-based loss** allows the network to focus on learning a meaningful and discriminative feature space by explicitly taking into account the links between triplets of examples. This is especially beneficial in applications that need fine-grained similarities or dissimilarities, such as facial recognition or picture retrieval. It consists of the Anchor A which is the reference instance for which we want to learn the representation. Positive P which is a similar instance to the anchor (same class or category). Negative N which is a dissimilar instance to the anchor (different class or category). Triplet-based loss can be calculates using the following formula in:

$$L(A, P, N) = max(0, d(A, P) - d(A, N) + \alpha) \quad \textbf{(3)}$$

- d (A, P) is the distance between the anchor and positive examples.
- d (A, N) is the distance between the anchor and negative examples.
- α is a margin or threshold that ensures the positive examples are sufficiently closer to the anchor than the negative examples.

FaceNet uses a deep convolutional neural network to generate a compact Euclidean space for face photos, allowing for fast face verification, identification, and grouping at scale. The model directly optimises the embedding with a triplet-based loss function, which consists of an anchor face, a matching positive face, and a non-matching negative face. The training procedure includes online triplet mining, which selects hard triplets to ensure efficient convergence. FaceNet uses harmonic embeddings to ensure compatibility across different network outputs and applies a harmonic triplet loss. The models are trained using Stochastic Gradient Descent and AdaGrad to investigate architectures with different parameters and FLOPS. The approach achieves cutting-edge performance on benchmark datasets, displaying resilience to posture and lighting fluctuations [4].

The FaceNet model uses a deep convolutional neural network (CNN) with two fundamental architectures: the Zeiler&Fergus model and the Inception model. The Zeiler&Fergus-inspired network has 22 layers and uses 1x1 convolutions between typical convolutional layers, yielding 140 million parameters and needing 1.6 billion floating-point operations per picture [5]. The Inception-based models, NNS1 and NNS2, have much fewer parameters (6.6M-7.5M) and fewer floating-point operations (220M-1.6B) such as in **Figure 2**. The network is trained with Stochastic Gradient Descent (SGD) and AdaGrad, and the triplet loss function is used to optimise a 128-dimensional embedding. The live triplet selection technique provides efficient triplet creation by focusing on difficult situations, which contributes to speedy model convergence.

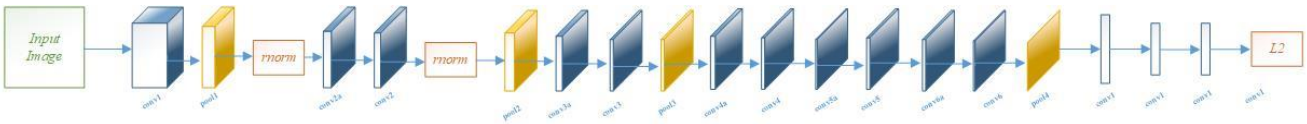


*Figure 2: FaceNet architecture. You can see the full details of the layer's sizes, The flow of the data and PARM, FLOPS in Table2*

| layer | size-in | size-out | kernel | param | FLPS |
|---|---|---|---|---|---|
| conv1 | 220×220×3 | 110×110×64 | 7×7×3, 2 | 9K | 115M |
| pool1 | 110×110×64 | 55×55×64 | 3×3×64, 2 | 0 | |
| rnorm1 | 55×55×64 | 55×55×64 | | 0 | |
| conv2a | 55×55×64 | 55×55×64 | 1×1×64, 1 | 4K | 13M |
| conv2 | 55×55×64 | 55×55×192 | 3×3×64, 1 | 111K | 335M |
| rnorm2 | 55×55×192 | 55×55×192 | | 0 | |
| pool2 | 55×55×192 | 28×28×192 | 3×3×192, 2 | 0 | |
| conv3a | 28×28×192 | 28×28×192 | 1×1×192, 1 | 37K | 29M |
| conv3 | 28×28×192 | 28×28×384 | 3×3×192, 1 | 664K | 521M |
| pool3 | 28×28×384 | 14×14×384 | 3×3×384, 2 | 0 | |
| conv4a | 14×14×384 | 14×14×384 | 1×1×384, 1 | 148K | 29M |
| conv4 | 14×14×384 | 14×14×256 | 3×3×384, 1 | 885K | 173M |
| conv5a | 14×14×256 | 14×14×256 | 1×1×256, 1 | 66K | 13M |
| conv5 | 14×14×256 | 14×14×256 | 3×3×256, 1 | 590K | 116M |
| conv6a | 14×14×256 | 14×14×256 | 1×1×256, 1 | 66K | 13M |
| conv6 | 14×14×256 | 14×14×256 | 3×3×256, 1 | 590K | 116M |
| pool4 | 14×14×256 | 7×7×256 | 3×3×256, 2 | 0 | |
| concat | 7×7×256 | 7×7×256 | | 0 | |
| fc1 | 7×7×256 | 1×32×128 | maxout p=2 | 103M | 103M |
| fc2 | 1×32×128 | 1×32×128 | maxout p=2 | 34M | 34M |
| fc7128 | 1×32×128 | 1×1×128 | | 524K | 0.5M |
| L2 | 1×1×128 | 1×1×128 | | 0 | |
| total | | | | 140M | 1.6B |

*Table 2: table show the structure of our Zeiler&Fergus based model with 1×1 convolutions. The input and output sizes are described in rows × cols × #filters. The kernel is specified as rows × cols, stride and the maxout pooling p=2* **[4]**.

ii. EfficientFace:

For emotion recognition we used EfficientFace, a lightweight and resilient facial expression recognition network designed for practical FER in the wild [6]. In the context of feature extraction, a local-feature extractor and a channel-spatial modulator are presented. The learnt face characteristics remain stable under occlusion and posture change.

To address the noise problem on facial expression datasets and the fact that real-world facial expressions are a distribution rather than a single emotion, a new LDL approach is presented based on LDG design. To reduce computational overheads in the FER model, we use the lightweight network ShuffleNet-V2, which includes Conv1, Stage2, Stage3, Stage4, and Conv5. To address occlusion and position variation in real-world scenarios, we created a local-feature extractor and channel-spatial modulator.

Furthermore, a new label distribution learning approach is given that is congruent with the psychologist's theory (Plutchik 1980) [7]. The next subsections introduce the proposed local-feature extractor and channel-spatial modulator, followed by information on the label distribution generator and loss.

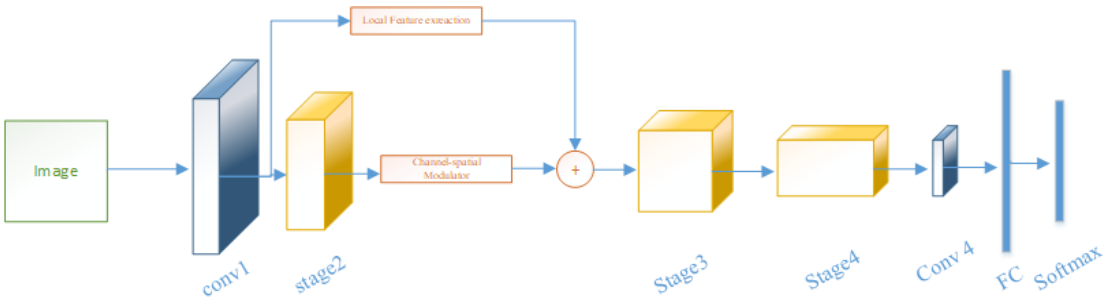


*Figure 3: Effiecientface abstract architecture showing its pipeline.*

A **local feature extractor** is a component that detects and extracts unique characteristics from various parts of an image. These characteristics are usually tiny, localised patterns or structures that are distinct or easily identifiable, such as corners, edges, or blobs [8]. The purpose is to collect information about significant spots or regions in a picture that may be used to characterise its content in a way that is unaffected by specific transformations such as translation, rotation, and scaling. The architecture of our local feature extractor is in **Figure4**.

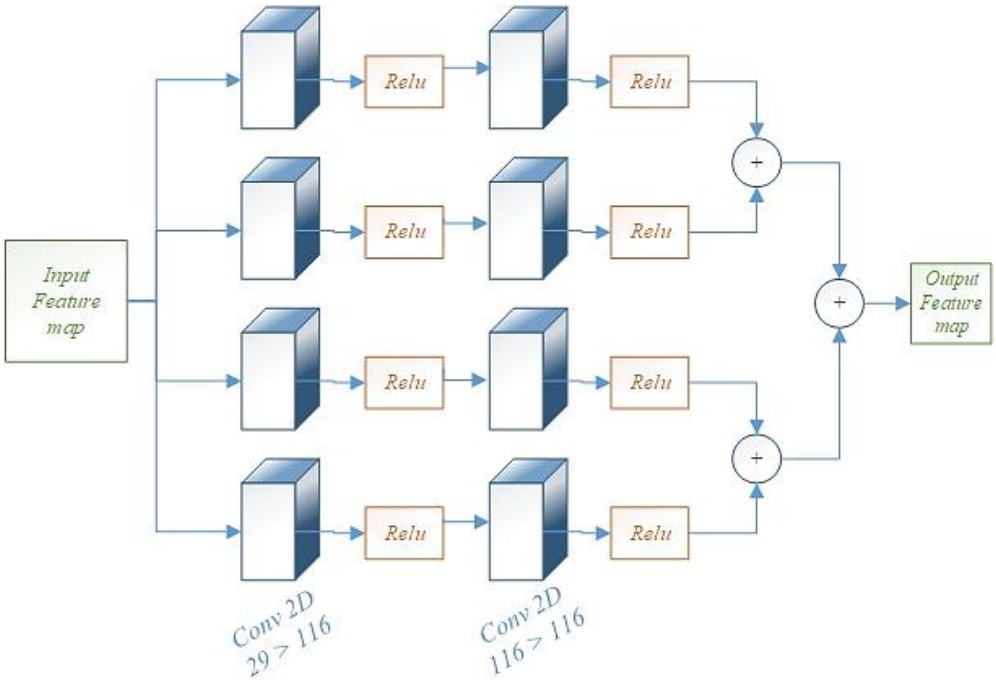


*Figure 4: Local-feature extractor used in EffecientFace components.*

Previous research suggests that learning local face traits is beneficial for FER in the wild. However, such approaches acquire local characteristics using face landmarks, which is inefficient. The model presents an efficient local-feature extractor that learns local area features at the feature level. These features are then supplemented into global features as residuals. The **channel-spatial modulator** is introduced to highlight the crucial global features after the Stage2. Allow for computational overheads, the channel-spatial modulator is designed based on BAM. The architecture of our local feature extractor is in **Figure5**.

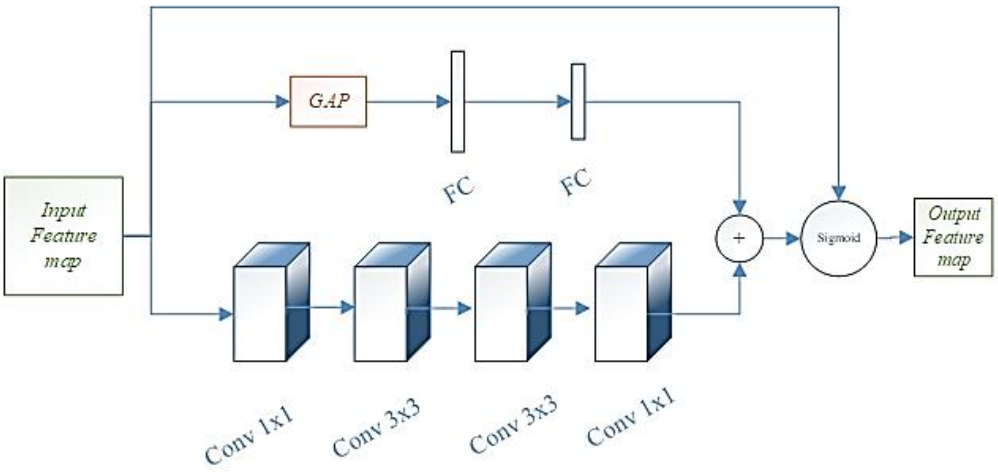


*Figure 5: Channel-spatial modulator compnents*

The model consists of several stages, each composed of **inverted residual** blocks. The stages are named 'stage2', 'stage3', and 'stage4', and their architecture is determined by the stages_repeats and stages_out_channels parameters provided during initialization. The InvertedResidual is a building block commonly used in mobile and efficient neural network architectures, particularly in models like MobileNetV2 and EfficientNet. This block is called "inverted" because it inverts the traditional residual block structure, where the shortcut connection is applied before the non-linear activation [9].

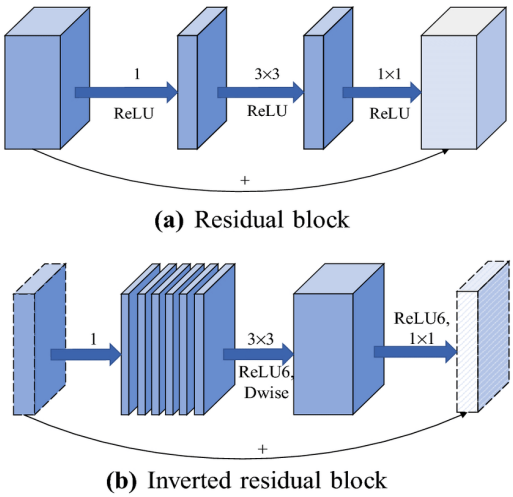


*Figure 6: (a) Represents the normal residual block which has a wide -> narrow -> wide structure with the number of channels. (b) Inverted Residual Block follows a narrow -> wide -> narrow approach, hence the inversion..*

**Audio Model**

For our audio model we used MMER (Multimodal Multi-task learning approach for Speech Emotion Recognition). MMER is based on early fusion and cross-modal self-attention between text and acoustic modalities and solves three auxiliary tasks for learning emotion recognition from spoken utterances. In practice, MMER outperforms all our baselines and achieves state-of-the-art performance on the IEMOCAP benchmark. Additionally, we conduct extensive ablation studies and results analysis to prove the effectiveness of our proposed approach [10].

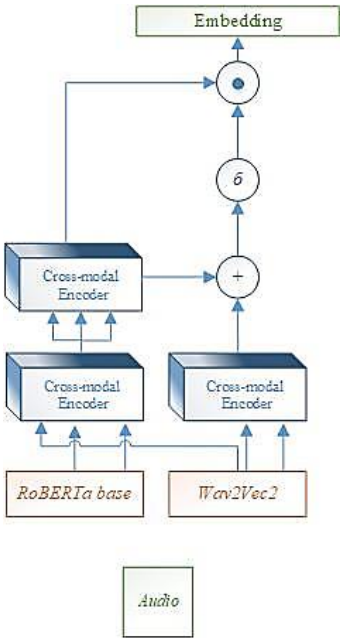


*Figure 7: MMER architecture*

For the pre-trained acoustic encoding it uses Wav2Vec2 which is a model made by Facebook. For the pre-trained text encoding it uses $\text{RoBERTa}_{\text{BASE}}$ which uses BERTs model. You can see the full architecture in **Figure7**.

The cross-modal encoder in Figure7 seeks to create a common representation space in which data from many modalities may be incorporated and compared. This shared representation captures important correlations and similarities between instances of data from many modalities. The fundamental objective is to develop a combined representation that allows for tasks requiring many types of data. You can see the details in **Figure8**.

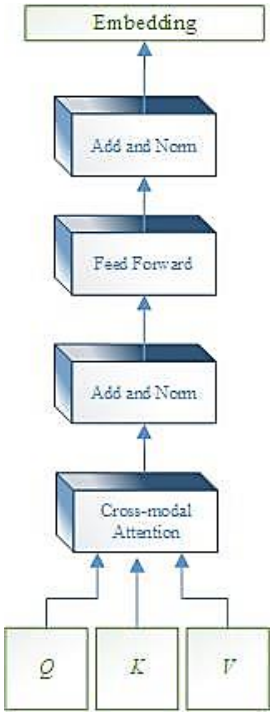


*Figure 8: Cross Modal Encoder*

### i. Wav2Vec

Wav2Vec2 is a speech model that accepts a float array corresponding to the raw waveform of the speech signal. wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task specified by a quantization of the latent representations that are simultaneously trained. Experiments employing all labelled Librispeech data provide 1.8/3.3 WER on clean/other test sets. When the quantity of labelled data is reduced to one hour, wav2vec 2.0 exceeds the prior state of the art on the 100-hour subset while using 100 times less labelled data.

Using only 10 minutes of labelled data and pre-training on 53k hours of unlabelled data yields 4.8/8.2 WER. This illustrates the capability of voice recognition with little labelled data. [11]. The model uses the pre-trained checkpoint released by Facebook, pre-trained on 960 hours of Librispeech, and use the wav2vec-2.0-base architecture for all our experiments [12]. You can see the pipeline of the model in **Figure9.**

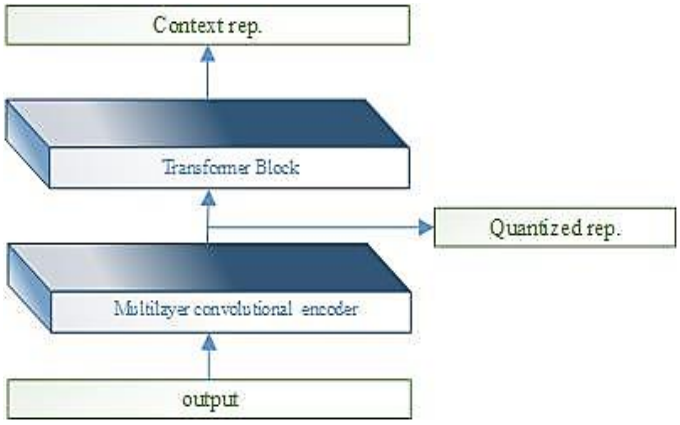


*Figure 9: Wav2Vec2 ver2.0 pipeline.*

### ii. RoBERTa$_{BASE}$

RoBERTa is a self-supervised transformers model trained on a vast corpus of English data. This implies that it was pretrained solely on raw texts, with no human labelling (which allows it to use a large amount of publicly available data), and that an artificial mechanism generated inputs and labels from those texts. RoBERTa iterates on BERT's pretraining procedure, including training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data [13].

# CHAPTER FOUR: SOLUTION ARCHITECTURE

The solution architecture for the Emotion Detection Deep Learning Model involves a comprehensive integration of preexisting visual and audio models to enhance accuracy. Fusion technique was used to enhance the performance of the overall model by fusing the two mentioned before models. We used the **Late** fusion as an approach to model fusion, where the predictions of individual models are combined at a later stage, typically after their outputs have been generated.

However, not only we used the outputs of each model directly to fuse the 2 models, but also, we experimented with different ways to apply the late fusion model that are discussed later in this chapter. But first let us talk about our project timeline.

## Project Timeline

### Visual Model (FaceNet and EfficientFace Integration)

In the early stages of the project, the integration process begins with the exploration of the FaceNet architecture. FaceNet, known for its advancements in face recognition, provides a solid foundation for extracting discriminative facial features from raw video data. The model's pre-trained weights and capabilities in generating face embeddings are crucial components in establishing a comprehensive visual representation for subsequent emotion recognition tasks.

Parallelly, the project delves into the integration of EfficientFace, a specialized architecture designed for efficient facial expression recognition directly from raw video frames. The decision to incorporate EfficientFace lies in its capacity to efficiently capture facial expressions, aligning with the project's emphasis on end-to-end learning without the need for separately learned features. This integration involves incorporating the EfficientFace feature extraction component into the visual branch of the overall audiovisual emotion recognition model.

As the integration progresses, the two models are harmoniously fused within the visual branch. The integration point involves strategically combining the extracted facial features from both FaceNet and EfficientFace, creating a rich and comprehensive representation of facial expressions. This fusion is crucial for capturing nuanced facial information that contributes to a more robust understanding of human emotions.

Throughout the integration process, considerations are made to optimize computational efficiency, ensuring that the final visual model performs effectively in real-world scenarios. The collaborative efforts of FaceNet and EfficientFace within the visual branch contribute to the model's ability to handle diverse facial expressions in unconstrained settings, addressing challenges where one modality may be absent or noisy during inference.

### Audio Model (MMER with Wav2Vec2 and RoBERTaBASE)

Now the part for the Audio Model within the context of Multi-Modal Emotion Recognition (MMER) involves the integration of advanced models, specifically Wav2Vec2 and RoBERTaBASE, to enhance the system's ability to analyse and recognize emotions from audio data. In this integrated approach, the Wav2Vec2 model is employed for audio feature extraction, capturing intricate patterns in speech signals, and converting them into meaningful representations. This process allows the model to directly learn from raw audio data, providing a foundation for end-to-end training without the need for pre-extracted features.

Complementing the audio feature extraction, the RoBERTaBASE model plays a crucial role in understanding the semantic context of the transcribed audio content. By leveraging RoBERTa's natural language processing capabilities, the model can capture nuanced linguistic cues, aiding in the interpretation of emotional states expressed through speech. The integration of RoBERTaBASE further enriches the overall multi-modal architecture by combining auditory information with contextual understanding derived from textual representations.

Throughout the project timeline, the focus lies on achieving seamless integration between the Wav2Vec2 and RoBERTaBASE models within the MMER framework. This integration not only enables the joint learning of audio and textual features but also ensures that the system remains robust in real-world scenarios where one modality might be incomplete or noisy. The proposed modality dropout mechanism, inspired by the challenges encountered in multi-modal emotion recognition, aims to enhance the model's resilience by simulating scenarios with missing or unreliable data during training. This holistic approach considers the interplay between audio and textual modalities, emphasizing not only performance gains in ideal settings but also the system's adaptability in less-than-ideal conditions.

## Fusion Implementation

In the pursuit of refining the Emotion Detection Deep Learning Model, the implementation of fusion strategies emerges as a critical component. Leveraging a late fusion approach, our objective is to seamlessly integrate information from both video and audio models. Late fusion, chosen for its efficacy and reduced training requirements, plays a pivotal role in aggregating diverse outputs into unified emotion predictions. Intermediate fusion was used for to achieve better performance using feature maps for both models.

This section delineates four distinctive fusion strategies each tailored to capture temporal patterns and optimize feature combination for enhanced emotion prediction. From straightforward concatenation to intricate architectures involving average pooling, dropout, and non-linear activations, these fusion methodologies are designed to unravel the complexity of human emotion.

Method 1

This method involves the consolidation of 16 outputs (8 output for each emotion from each model), corresponding to different emotions. These diverse outputs are aggregated to produce 8 final predictions. This straightforward fusion method sets the foundation for the subsequent strategies, providing a baseline for comparison with more intricate architectures.

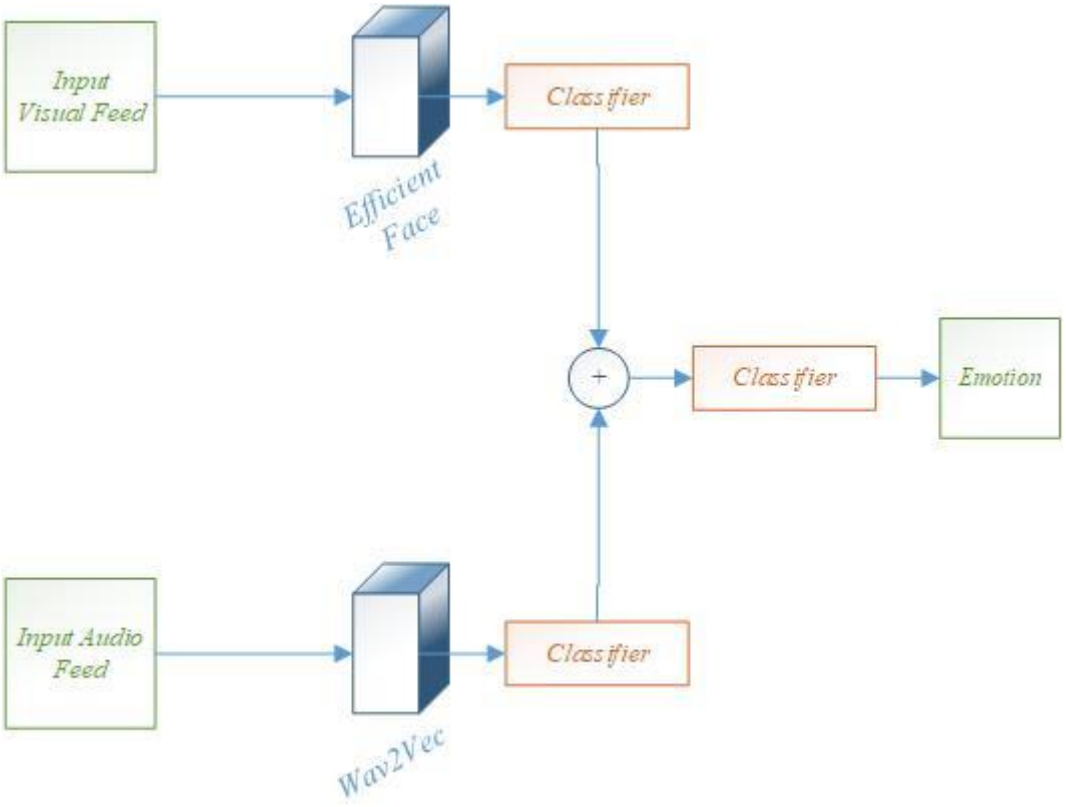


**Figure 10 Fusion 1**

**CHANGES:**

The individual models were not changed.

**FUSION:**

- There are 16 outputs in total, 8 from each model corresponding to different emotions.
- The fusion involves aggregating these 16 outputs into 8 final predictions using 1 additional layer.
- Demonstrates high validation accuracy (95.83%) and solid test accuracy (93.75%).

Method 2

For the second method used, a more complex architecture unfolds. The fusion involves concatenating output representations from both models, also the last layer from the visual model is neglected and the output is received from the previous layer and the same is done for the audio model. Furthermore, a linear layer is employed to reduce the concatenated dimension to 8 outputs for each emotion.

This method allows us to have a broader view of the features capture by each model instead of just using the final output of each model such as the previous model.

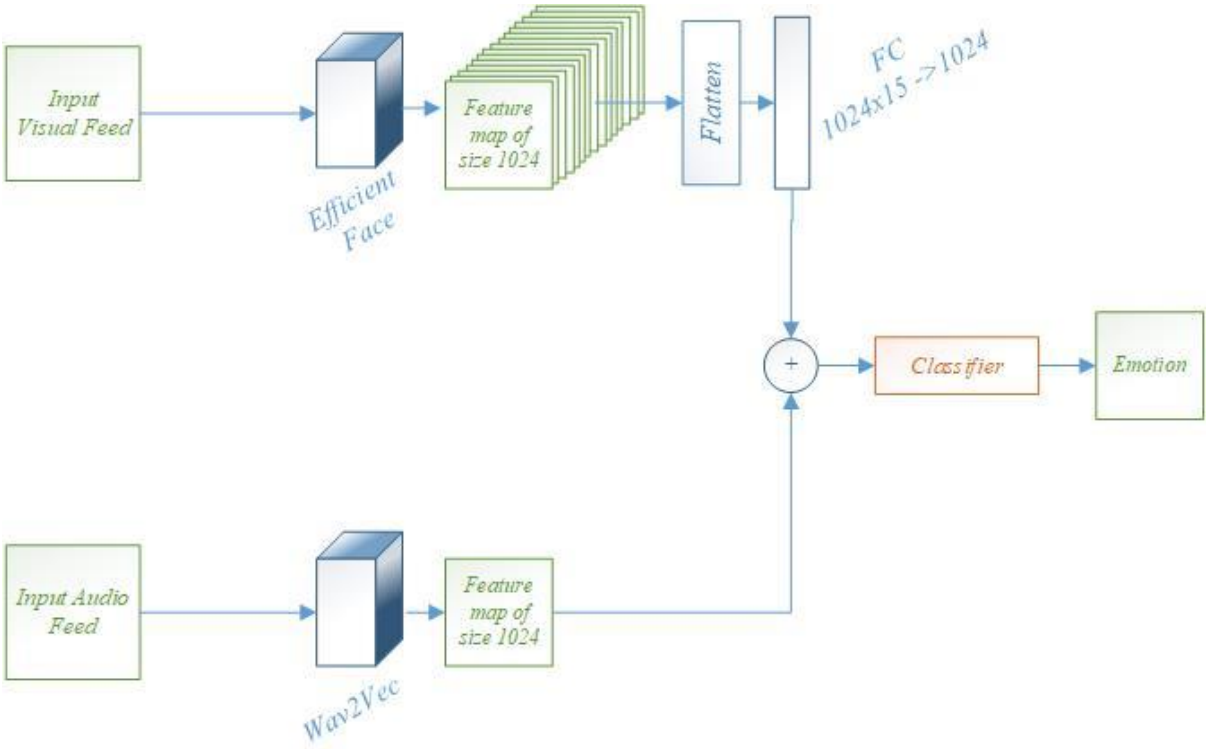


**Figure 11 Fusion 2 (removed last layer of each model)**

**CHANGES:**

- Remove last layer from both the audio and video models.
- Added a fully connected layer to the video model to map the (1024*15) to 1024 feature.
- The audio model now outputs 1024 since the last layer is removed

**FUSION:**

- The fusion involves aggregating these 2048 (1024 output from each model) outputs
- A linear layer is used to reduce the concatenated dimension from 1024*2 to 8.
- Although exhibiting lower validation (81.5%) and test (74.38%) accuracies, lays the groundwork for subsequent, more intricate fusion methods.

Method 3

Here, a similar architecture to the previous method is employed, with the concatenation of output representations from visual and audio models. However, instead of a linear layer, average pooling is applied along the temporal dimension, followed by a subsequent linear layer, capturing temporal patterns in a different manner.

This method helped in reducing the training time, since average pooling helped to reduce our number of captured features and this layer have no learnable parameters so training is not heavily affected by the introduction of this layer.

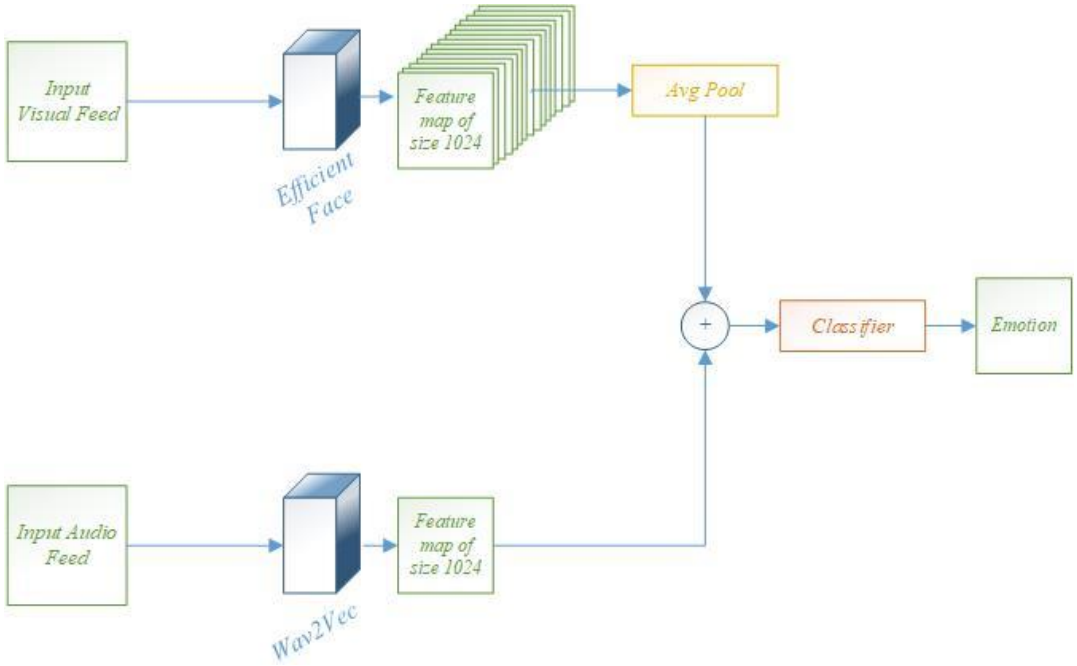


**Figure 12 Fusion 3 (Used Average Pool instead of FC in visual model)**

**CHANGES:**

- Remove last layer from both the audio and video models.
- Added an average pooling layer to the video model to map the (1024*15) to 1024 feature.
- The audio model outputs 1024 since the last layer is removed

**FUSION:**

- The fusion involves aggregating these 2048 (1024 output from each model) outputs
- A linear layer is used to reduce the concatenated dimension from 1024*2 to 8.
- Achieves impressive validation (97.08%) and test (96.04%) accuracies, showcasing the efficacy of average pooling in reducing training time without compromising performance.

Method 4

In this approach, we enhance the fusion strategy by introducing more complexity. We add extra layers, including dropout for regularization and an ELU activation function. These modifications aim to better capture the emotional cues by emphasizing non-linear activations and implementing regularization techniques.

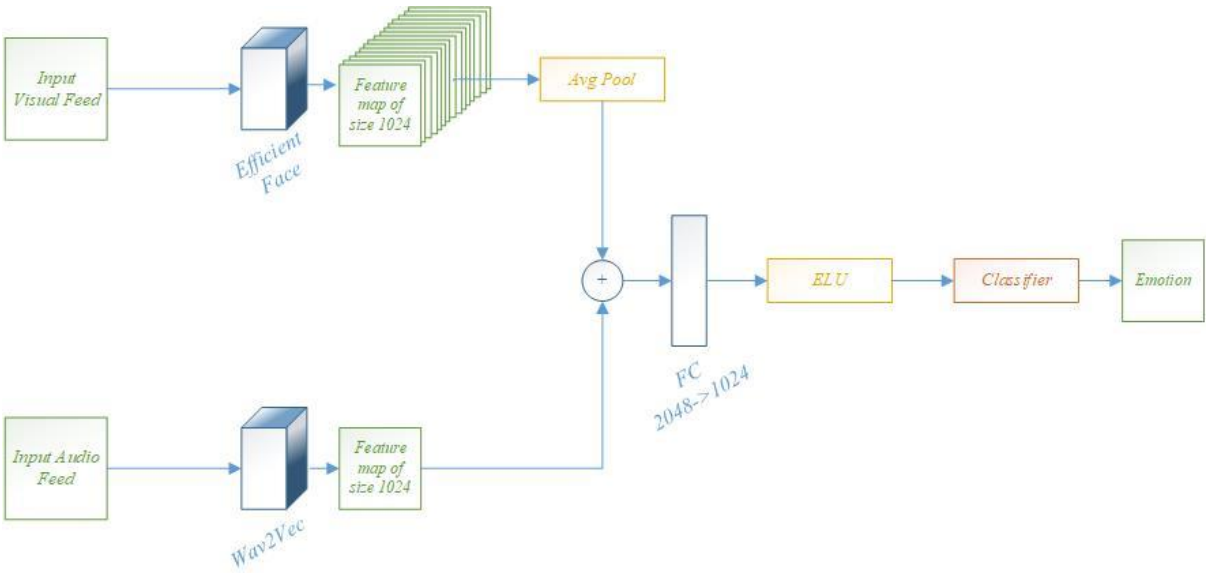


**Figure 13 Fusion 4 (Using ELU and Dropout)**

**CHANGES:**

- Remove last layer from both the audio and video models.
- Added an average pooling layer to the video model to map the (1024*15) to 1024 feature.
- The audio model outputs 1024 since the last layer is removed

**FUSION:**

- The fusion involves aggregating these 2048 (1024 output from each model) outputs
- A linear layer is used to reduce the concatenated dimension from 1024*2 to 1024.
- A dropout layer is introduced for regularization.
- An ELU activation function is applied.
- A linear layer reduces the dimension from 1024 to 8 for the final predictions.
- Introduces additional complexities, leading to excellent validation (97.58%) and test (97.08%) accuracies, emphasizing the significance of non-linear activations and regularization techniques.

Method 5

Similar to the previous method, adds complexity by incorporating additional layers, including dropout for regularization and an ELU activation function. The key distinction is the use of max pooling instead of average pooling, offering an alternative approach to capturing temporal patterns in the fused feature space.

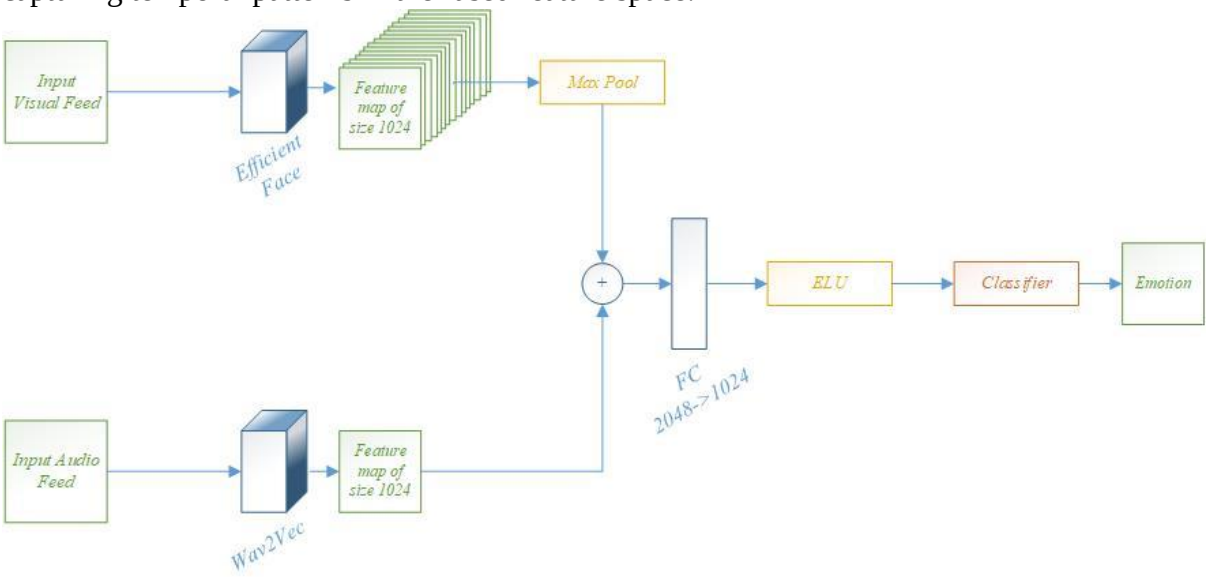


**Figure 14 Fusion 5 (Using ELU, Dropout, and Max Pooling)**

**CHANGES:**

- Remove the last layer from both the audio and video models.
- Add an average pooling layer to the video model to map the (1024*15) to 1024 features.
- The audio model outputs 1024 since the last layer is removed.

**FUSION:**

- Aggregate the 2048 outputs (1024 from each model).
- Apply a linear layer to reduce the concatenated dimension from 2048 to 1024.
- Introduce a dropout layer for regularization.
- Apply an ELU activation function.
- Utilize max pooling instead of average pooling to capture diverse temporal patterns.
- Conclude with a linear layer to reduce the dimension from 1024 to 8 for the final predictions.
- With max pooling, maintains high validation (97.92%) and test (97.08%) accuracies, illustrating the effectiveness of incorporating max pooling for capturing diverse temporal patterns.

# CHAPTER FIVE: SOFTWARE REQUIREMENTS SPECIFICATIONS (SRS)

## Scope

The proposed project attempts to create an Emotion Recognition Deep Learning Model that uses visual and auditory inputs to increase accuracy. The system consists of two separate models: one for processing facial image sequences using an efficient face model, and another for analysing audio data using the Wave2Vec2 model. These distinct models will be trained using the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song).

The system will be designed to operate in an environment that supports the necessary deep learning frameworks, libraries, and dependencies for both image and audio processing.

It is assumed that the RAVDESS dataset adequately represents a diverse range of emotional expressions and can be used effectively for training both the visual and auditory models.

Dependencies include the availability of libraries and frameworks compatible with the chosen models and the accessibility of required computational resources.

**Resource Constraints:** The project will be constrained by available computational resources, including GPU capabilities, and will need to optimize model architecture accordingly.

**Time Constraints:** The project is bound by a timeline corresponding to the academic schedule for the graduation project.

The main aspects of the Emotion Recognition Deep Learning Model are:

- Efficient Face Model: An efficient face model is used to analyze facial image sequences in order to extract emotional information.
- Wave2Vec2 Audio Model: Using the Wave2Vec2 model to process audio data and extract emotional clues from speech and music.
- Training on RAVDESS: The efficient face model and the audio model will be trained using the RAVDESS dataset, which contains a wide range of emotional expressions.
- Fusion of models: The combination of visual and aural models to generate a final Emotion Detection Model, which improves accuracy by incorporating input from both modalities.

## Functional Requirements

### Data Input

- Image Input: The system should accept sequences of facial images as input for emotion prediction.
- Audio Input: The system should accept audio data as input for emotion prediction, where the audio data corresponds to speech or song.

### Model Training

- Efficient Face Model: The system should train the efficient face model on the RAVDESS dataset, ensuring the extraction of facial features relevant to emotion recognition.
- Wave2Vec2 Audio Model: The system should train the Wave2Vec2 model on the RAVDESS dataset, capturing emotional cues from the audio data, including speech and song.

### Fusion of Models

- Integration: The system should integrate the outputs of the trained efficient face model and the Wave2Vec2 audio model to create a fused Emotion Detection Model.
- Fusion Technique: The system should employ a suitable fusion technique, such as concatenation or ensemble methods, to combine the predictions from the visual and auditory models.

### Prediction

- Emotion Prediction: The system should provide the capability to predict emotions based on new input data, using the final Emotion Detection Model.
- Output Format: The predictions should be presented in a structured format, indicating the likelihood or probability of each emotion category.

### User Interfaces

- Training Configuration Interface: The system should provide a user interface for configuring training parameters, including options for adjusting hyperparameters, selecting training epochs, and specifying batch sizes.
- Monitoring Interface: The system should offer a monitoring interface for users to track the progress of the training process, visualize training metrics, and identify potential issues.
- Interaction Interface: The system should include an interface for users to interact with the final Emotion Detection Model, providing input for prediction and receiving the model's output.

### Hardware Interfaces

- Compatibility: The system should be compatible with standard hardware configurations, and the efficient utilization of GPU resources should be considered for model training.

### Software Interfaces

- Deep Learning Frameworks: The system should interface with deep learning frameworks, including but not limited to TensorFlow or PyTorch, for model development, training, and deployment.
- Data Storage: The system should interact with data storage systems for managing and accessing the RAVDESS dataset during training.

## Non-Functional Requirements

#### Performance Requirements

- Training Time: The training process for both the efficient face model and the Wave2Vec2 audio model should be completed within a reasonable time frame, considering computational resources.
- Inference Time: The Emotion Detection Model should provide predictions for new input data in real-time or within acceptable response time limits.

#### Reliability

- Robustness: The Emotion Detection Model should handle variations in input data gracefully, providing consistent and reliable predictions across different scenarios.

#### Scalability

- Model Enhancement: The system should be designed to accommodate future enhancements, including the integration of additional modalities or improvements to existing models.
- Dataset Variations: The Emotion Detection Model should demonstrate scalability in handling variations in input data, such as diverse facial expressions and emotional nuances in speech and song.

## Use Cases

**1. Train Efficient Face Model**

*Primary Actor:* Machine Learning Engineer

*Description:* The machine learning engineer starts the training procedure for the efficient face model. The use case consists of loading the RAVDESS dataset, establishing training parameters (such as batch size and learning rate), and running the training script. The system should provide feedback on training progress, show important metrics, and save the learned model for future use.

**2. Train Wave2Vec2 Audio Model**

*Primary Actor:* Machine Learning Engineer

*Description:* The machine learning engineer begins training the Wave2Vec2 audio model. This includes loading the RAVDESS audio dataset, configuring the training settings, and running the training script. The system should give real-time feedback on training progress, show relevant metrics, and preserve the trained audio model.

**3. Fuse Models for Emotion Detection**

*Primary Actor:* System

*Description:* To generate the final Emotion Detection Model, the system combines the trained efficient face model with the Wave2Vec2 audio model. This entails using a fusion technique (such as concatenation) to integrate the visual and audio model outputs. The fused model is then retained for use in future emotion prediction challenges.

**4. Predict Emotions from Images and Audio**

*Primary Actor:* End User

*Description:* The end user interacts with the Emotion Detection Model, which predicts emotions based on input data. The user inputs a sequence of facial photos or audio data, which the system processes using trained models. The result is a structured list of projected emotions, along with related likelihood or probability scores.

**Deliverables with estimated time plan (Semester 1)**

- **-Familiarize** ourselves with existing research and techniques related to visual emotion detection.
  **-Identify** potential datasets for visual emotion detection and begin acquiring them. (Weeks 1-2)

- **-Preprocess** and clean the acquired visual emotion datasets.
  **-Explore** different computer vision techniques and models suitable for emotion detection.
  **-Implement** and train a visual emotion detection model using one of the selected datasets.
  (Weeks 3-4)

- **-Evaluate** and fine-tune the performance of the visual emotion detection model.
  **-Analyze** the results and iterate on the model if necessary.
  **-Prepare** interim progress reports and documentation.
  (Weeks 5-6)

- **-Conduct** additional experiments or refine the model based on feedback and evaluation.
  **-Optimize** the visual emotion detection model for real-time or near-real-time inference.
  **-Validate** the performance of the visual emotion detection model using separate test datasets.
  (Weeks 7-9)

- **-Document** the methodology, findings, and limitations of the visual emotion detection model.
  **-Prepare** a presentation summarizing the work accomplished in the first semester.
  **-Study** speech emotion detection algorithms and techniques.
  **-Begin** acquiring relevant speech emotion datasets.
  (Weeks 10-13)

**5. Configure Training Parameters**

*Primary Actor:* Machine Learning Engineer

*Description:* The machine learning engineer uses the training configuration interface to specify settings for training the efficient face model and the Wave2Vec2 audio model. This includes modifying hyperparameters, choosing the number of training epochs, and determining batch sizes. The system checks the input parameters for accuracy and consistency.

**6. Monitor Training Process**

*Primary Actor:* Machine Learning Engineer

*Description:* The machine learning engineer utilizes the monitoring interface to track the training progress of both the efficient face model and the Wave2Vec2 audio model. The interface shows real-time indicators like loss and accuracy, allowing the user to recognize possible problems, make informed decisions, and modify the training process as needed.

**7. Interact with Emotion Detection Model**

*Primary Actor:* End User

*Description:* The end user communicates with the Emotion Detection Model via the interaction interface. This comprises supplying input data for emotion prediction, obtaining predictions, and analyzing the model's results. The interface should be user-friendly and allow for easy contact with the model.

**8. Handle Unexpected Input Gracefully**

*Primary Actor:* System

*Description:* The system ensures robustness by gently processing unexpected input data during the prediction process. If the input data deviates from the intended format or contains anomalies, the system should handle the errors and offer relevant feedback to the user.

# CHAPTER SIX: PROTOTYPE RESULTS

Our model has yielded promising results across various fusion strategies. When examining individual models, Efficient Face demonstrated a testing accuracy of 74.92%, while Wav2Vec ver2 outperformed with a testing accuracy of 91.25%. The combination of both modalities in Fusion 1 showcased a substantial improvement, achieving a high testing accuracy of 93.75%, suggesting effective fusion of facial and audio features.

Among the explored fusion strategies, Fusion 3 emerged as particularly successful, employing a combination of 1024-dimensional features from both modalities using average pooling. This fusion approach demonstrated outstanding accuracy in both validation (97.08%) and testing (96.04%) datasets. Moreover, deeper fusion strategies, as seen in Fusion 4 and Fusion 5, involving additional pooling layers, exhibited even higher accuracy, reaching up to 97.92% in validation and 97.08% in testing. In contrast, Fusion 2, which utilized fully connected layers for fusion, showed comparatively lower accuracy, highlighting the importance of choosing an effective fusion strategy for multimodal models.

| Model | Fusion Type | Validation Accuracy | Testing Accuracy |
|---|---|---|---|
| Efficient Face | NA | - | 74.92% |
| Wav2Vec ver2 | NA | - | 91.25% |
| Our Model | Fusion 1 (8+8 -> 8) | 95.83% | 93.75% |
| | Fusion 2 (1024+1024 -> 8 using FC) | 81.50% | 74.38% |
| | Fusion 3 (1024+1024 -> 8 using Avg Pool) | 97.08% | 96.04% |
| | Fusion 4 (1024+1024 -> 1024 -> 8 using Avg Pool) | 97.58% | 97.08% |
| | Fusion 5 (1024+1024 -> 1024 -> 8 using Max Pool) | 97.92% | 97.08% |

*Table 3: Performance of different fusion methods on RAVDESS DB compared to the original models.*

In summary, the multimodal emotion recognition model's success lies in the strategic fusion of facial and audio features. Fusion 3, Fusion 4, and Fusion 5, characterized by pooling operations, showcased superior accuracy in recognizing emotions, indicating the potential for more sophisticated fusion strategies in multimodal models. These findings contribute valuable insights for optimizing multimodal architectures for emotion recognition tasks.

# CHAPTER SEVEN: CONCLUSIONS

In this project, we extensively investigated and integrated sophisticated neural network architectures for multimodal emotion identification. The visual model used FaceNet to recognise faces and EfficientFace to recognise facial expressions, whilst the audio model used Wav2Vec2 to recognise voice emotions. The fusion of these modalities was accomplished by late fusion, which combined the strengths of visual and aural information to improve emotion identification.

The visual model proved the usefulness of FaceNet in face identification by using a triplet-based loss function for optimum embedding in a small Euclidean space. EfficientFace, a lightweight and durable facial expression recognition network, tackled real-world problems by integrating local-feature extraction and channel-spatial modulation. Wav2Vec2 is an audio model used for acoustic feature extraction.

The fusion of visual and audio modalities was explored through various strategies. Fusion 3, involving average pooling of 1024-dimensional features, emerged as particularly successful with outstanding accuracy in both validation (97.08%) and testing (96.04%) datasets. Deeper fusion strategies in Fusion 4 and Fusion 5, incorporating additional pooling layers, exhibited even higher accuracy, reaching up to 97.92% in validation and 97.08% in testing. These results emphasize the importance of strategic fusion methods in achieving superior multimodal emotion recognition.

The project's timeline showcased the integration process of FaceNet and EfficientFace within the visual model, highlighting the harmonious fusion of facial features. Similarly, the integration of Wav2Vec2 and RoBERTaBASE within the MMER framework enriched the audio model's ability to understand emotion expressed through speech. The fusion implementation involved late fusion techniques, and the comparison of different fusion methods demonstrated the superiority of strategies emphasizing pooling operations.

To summarise, this effort proposes a strong multimodal emotion identification model that efficiently combines visual and aural data. The great accuracy in emotion detection achieved using multiple fusion techniques demonstrates the promise for real-world applications in understanding and interpreting human emotions. The findings provide useful insights for optimising multimodal systems and pave the path for breakthroughs in emotion identification technologies.

## REFERENCES

[1] B. H. M. H. M. Ali Mollahosseini, AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild, arvix.org, 2017.

[2] P. T. C. W. A. S. Y. T. P. I. A. M. C. L. D. K. Prannay Khosla, Supervised Contrastive Learning, arxiv.org, 2021.

[3] H. Wang, X. Li, Z. Ren, M. Wang and C. Ma, Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion., 2023.

[4] S. &. A. A. &. M. M. &. D. S. Boulahia, Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition, 2021.

[5] U. T. S. R. H. S. D. M. Sreyan Ghosh, MMER: Multimodal Multi-task Learning for Speech Emotion Recognition, arxiv.org, 2022.

[6] J. B. L. S. Kilian Q. Weinberger, Distance metric learning for large margin nearest neighbor classification, MIT Press, 2006.

[7] D. K. J. P. Florian Schroff, FaceNet: A Unified Embedding for Face Recognition and Clustering, arxiv.org, 2015.

[8] R. F. Matthew D Zeiler, Visualizing and understanding Convolutional Networks, arxiv.org, 2013.

[9] H. Kim, J.-H. Lee and B. C. Ko, Facial Expression Recognition in the Wild Using Face Graph and Attention, ieeexplore.ieee.org, 2023.

[10] Q. L. F. Z. Zengqun Zhao, Robust Lightweight Facial Expression Recognition Network with Label Distribution Training, aaai.org, 2021.

[11] П. С. Anatolij Sergiyenko, Local Feature extraction in images, http://itvisnyk.kpi.ua/, 2021.

[12] A. H. M. Z. A. Z. L.-C. C. Mark Sandler, MobileNetV2: Inverted Residuals and Linear Bottleneck, arxiv.org, 2018.

[13] H. Z. A. M. M. A. Alexei Baevski, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, arxiv.org, 2020.

[14] U. T. S. R. H. S. D. M. Sreyan Ghosh, MMER: Multimodal Multi-task Learning for Speech Emotion Recognition, 2023: arxiv.org.

[15] M. N. J. M. D. O. M. L. V. YinhanLiu, RoBERTa:ARobustlyOptimizedBERTPretrainingApproach, arxiv.org, 2019.