



Ain Shams University
Faculty of Computer & Information Sciences
Computer Science Department

Audio Mood Visualizer

(Automatic speech emotion visualization through avatars)

By:

Omnia Ahmed Elsaid Mostafa
Alaa Salah Abd Elhady Ragab
Amany Elsayed Mohamed Mohamed
Nour Elhoda Ahmed
Nourhan Adel Abd Elrady
Mai Adel Abdelaziz

Under Supervision of:

Dr. Hanan Hindy [BSc, MSc, PhD],
Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

A.L. Yomna Ahmed [BSc, MSc],
Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

July 2024



Ain Shams University
Faculty of Computer & Information Sciences
Computer Science Department



July 2024

Acknowledgement

We are deeply grateful to God for His unwavering guidance and support throughout this project, which has enabled us to overcome challenges and achieve our goals. His presence has been a constant source of strength, guiding our efforts and illuminating our path.

We extend our heartfelt gratitude to all who supported us in completing this project, especially our supervisors Dr. Hanan Hindy and A.L. Yomna Ahmed for their invaluable guidance and ideas.

Special thanks go to our parents and friends whose encouragement and suggestions were indispensable in bringing this project to fruition.

Abstract

“When dealing with people, remember you are not dealing with creatures of logic, but creatures of emotions.”

Dale Carnegie

Emotions constitute a fundamental aspect of human cognition, profoundly influencing interpersonal interactions. However, challenges such as social phobia can impede individuals from effectively expressing their emotions during communication.

Our Objective in this project is to create an application that makes it easier for people to communicate with each other and show their emotions. We aimed to achieve this by visualizing the emotions of someone talking -using audio signals only- through the generation of an appropriate avatar that reflects the emotion detected from speech. This can be achieved through applying recent advanced techniques in Speech Emotion Recognition.

Our framework can be divided into two main sub modules: Speech Emotion Recognition and Avatar Generation. In order to reach the best speech emotion recognition model, we applied extensive experiments using different models and datasets till the best combination was achieved. This combination was using the SER with CNN and Multi-head Convolutional Transformer model trained on the RAVADESS dataset which achieved 95% accuracy. This framework aims to facilitate emotional expression for individuals who may struggle with verbal communication, potentially revolutionizing interpersonal interactions and applications in various fields. Finally, our project focuses on enhancing human communication by visualizing emotions through avatars generated from speech.

Table of Contents

Acknowledgement	i
Abstract.....	ii
List of Figures	v
List of Tables	vi
List of Abbreviations	vii
1- Introduction	1
1.1 Motivation.....	1
1.2 Problem Definition.....	1
1.3 Objectives.....	2
1.4 Time Plan.....	3
1.5 Document Organization	3
2- Literature Overview	5
2.1 Project Overview	5
2.2 Theoretical Background.....	5
2.3 Related Work	7
2.3.1 Speech Emotion Recognition	8
2.3.2 Avatar Generation.....	14
-3 System Architecture and Methods	17
3.1 System Architecture.....	17
3.2 Description of Methods and Procedures	18
3.2.1 Presentation Layer	18
3.2.2 Logic Layer.....	18
3.2.2.1 Speech emotion recognition.....	18
3.2.2.2 Avatar Generation	20
3.2.3 Data Layer	22
3.3 System Users.....	22
4- System Implementation and Results	23
4.1 Datasets	23
4.2 Software Tools Used	25
4.3 Setup Configuration (software)	26
4.4 Experiments and Results.....	28
4.4.1 Speech emotion recognition experiments.....	28

4.4.2 Avatar Generation experiments.....	38
4.4.3 Realtime	40
5- User Manual.....	41
6- Conclusion and Future Work	46
6.1 Conclusion.....	46
6.2 Future Work.....	47
References	48

List of Figures

Figure 1.1: Time Plan	3
Figure 2.1: Two parallel CNNs with Transformer encoder for feature extraction. The extracted features are fed to the dense layer with a log SoftMax classifier for emotional state prediction.	11
Figure 2.2: The attention-guided generators process [10]	15
Figure 3.1: System Architecture	17
Figure 3.2: Panoramic view of the SER model used.	18
Figure 3.3: Two parallel CNN	19
Figure 3.4: Transformer-Encoder.	20
Figure 4.1: Loss curve for LSTM model [15].....	30
Figure 4.2: Confusion matrix for LSTM model [15].....	30
Figure 4.3: Loss Curve for CNN and Multi-Head Convolutional Transformer model [8].....	33
Figure 4.4: Confusion matrix for CNN and Multi-Head Convolutional Transformer model [8]on RAVDESS [12] dataset.....	33
Figure 4.5: Confusion matrix for CNN and Multi-Head Convolutional Transformer model [8]on SAVEE [13] dataset.	36
Figure 4.6: Confusion matrix for CNN and Multi-Head Convolutional Transformer model [8]on RAVDESS [12] + SAVEE [13] dataset.	37
Figure 4.7: Calm Figure 4.8: Happy Figure 4.9: Disgust	39
Figure 4.10: Sad Figure 4.11: Fearful Figure 4.12: Surprise.....	39
Figure 4.13: Angry.....	39
Figure 4.14: Fear with opened mouth.	40
Figure 4.15: Fear with closed mouth.	40
Figure 5.1: Home page.....	41
Figure 5.2: User Registration.....	42
Figure 5.3: User Login.....	42
Figure 5.4: Choose female avatar	43
Figure 5.5: Choose male avatar	43
Figure 5.6: Upload audio page.....	44
Figure 5.7: Allows users to choose audio file from their device to upload	44
Figure 5.8: Record their voice using the app	44
Figure 5.9: Waiting for creation of the video	45
Figure 5.10: Show video	45

List of Tables

Table 2.1: Results of SepTr and its ablated versions in comparison with various state-of-the-art methods on CREMA-D [7]	10
Table 2.2: Results of SepTr versus various state-of-the-art methods on Speech Commands V2 (SCV2) and ESC-50 [7].	10
Table 2.3: Summarized Overview of Significant SER Papers	13
Table 2.4: Quantitative comparison with different models. For all metrics except MSE, higher is better2.4 [10]	16
Table 4.1: Used datasets comparison.	25
Table4.2: Architecture changes in LSTM model [15]	32
Table4.3: Preprocessing changes in CNN and Multi-Head Convolutional Transformer model [8].	34
Table4.4: Architecture changes in CNN and Multi-Head Convolutional Transformer model [8].	35
Table 4.5: Datasets trials in CNN and Multi-Head Convolutional Transformer model [8].	35

List of Abbreviations

Abbreviation	Description
API	Application programming interface
BLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CPU	Central processing unit
DT	Decision Tree
DCNNs	Deep Convolution Neural Networks
DCT	discrete cosines transform
GMM	Gaussian mixture model
GANs	Generative Adversarial Networks
GPU	Graphics processing unit
HMM	Hidden Markov model
KNN	K-Nearest Neighbors
KR	Kernel Regression
LSTM	Long Short-Term Memory
MLB	Maximum Likelihood Bayes
MFCC	Mel-Frequency Cepstral Coefficients
NN	Neural Network
PNG	Portable Network Graphic
PDF	portable document format
RNN	Recurrent Neural Network
RBF	Radial Basis Function
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
SER	Speech Emotion Recognition
SVM	Support Vector Machine
SepTr	Separable Transformer
SAVEE	Surrey Audio-Visual Expressed Emotion
SVG	Scalable vector graphics
TESS	Toronto emotional speech set
VAEs	Variational Auto-Encoders

1- Introduction

1.1 Motivation

Our main motivation was to train an efficient model that detects emotions from voice and visualizes these emotions through an avatar. Such an application can be utilized in many fields:

- Medical field: can help people who have social phobias to communicate easily with others.
- Games: Deliver a virtual reflection of a player's emotions while protecting their privacy.
- Customer service: can aid the call center workers.
- E-learning and virtual meetings: help teachers to detect the emotional state of the students throughout the sessions.
- Supports privacy: You can use a photo that isn't yours or by using a default avatar.

1.2 Problem Definition

Our aim in this project is to develop a system capable of accurately detecting and classifying human emotions from audio signals. This entails tasks such as emotion recognition, classification, feature extraction, model training, and real-time

processing. However, several challenges complicate this endeavor. Emotions are inherently subjective and ambiguous, making it difficult to define clear boundaries between different emotional states. Obtaining accurate labels for emotional content in audio datasets is challenging due to the subjective nature of emotions and the lack of consistent ground truth. Moreover, emotions are often expressed through multiple modalities, leading to limited context and incomplete understanding when extracting emotions solely from audio. Addressing these challenges requires interdisciplinary research and careful consideration of privacy concerns, ethical implications, and the cultural variability in emotion expression.

1.3 Objectives

Our main objectives can be summarized as follows:

- 1- Apply recent advances in Speech Emotion Recognition (SER) research.
- 2- Visualize the recognized emotions (Happy, Sad, Fearful, Calm, Angry) through the generation of avatars with facial expressions that correspond to the detected emotion.
- 3- Create an application where the user could communicate with people and all his/her emotions are delivered to them through an avatar without needing to open the camera.
- 4- Experiment with additional features to further personalize the avatars while preserving the identity of the original speaker.

1.4 Time Plan

Figure 1.1 shows the overall project stages from learning until modules integration wherein work on this project was divided into several phases. First, we extensively surveyed relevant research and learnt about our project's domain. Second, we specified the requirements and set the system architecture and design, then we started on our main modules: SER and avatar generation. Finally, project testing and modules integration were applied. Work on the documentation was an ongoing process.

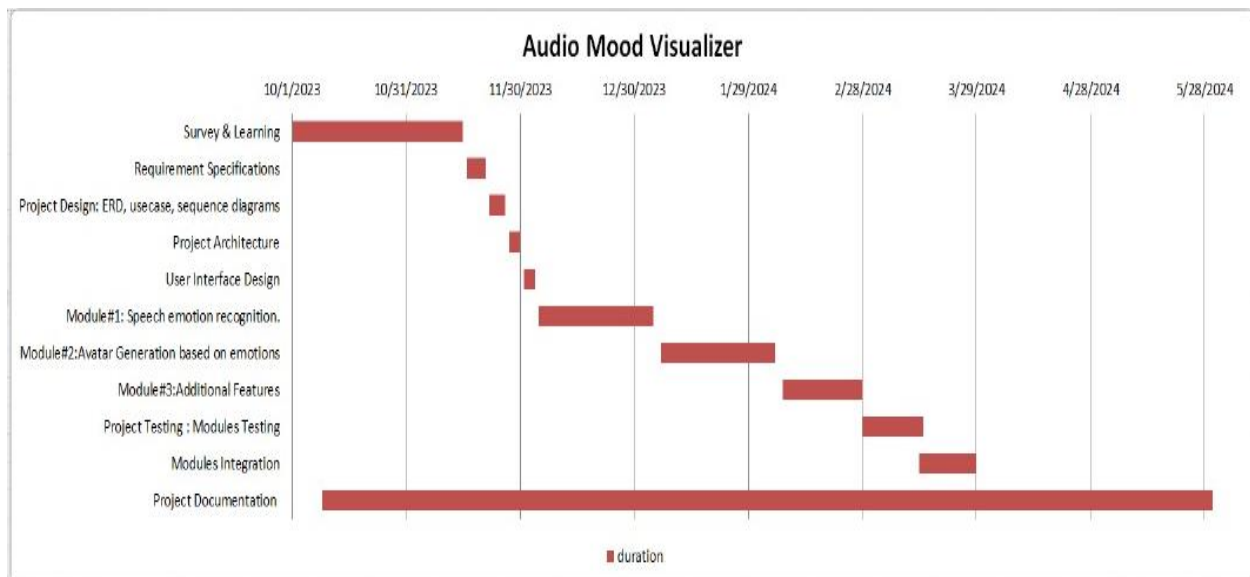


Figure 1.1: Time Plan

1.5 Document Organization

- **Chapter 2:** This chapter discusses the literature overview. It is divided into project overview, theoretical background and related works about speech emotion recognition and avatar generation.
- **Chapter 3:** This chapter discusses the three-tier system architecture with description to each of its components and the architecture of the used deep learning models.

- **Chapter 4:** This chapter explains datasets used to train the SER models, software tools used, setup configuration, and experiments and results that were applied with the models
- **Chapter 5:** This chapter outlines the detailed user's manual on how to use the system with a step-by-step screenshot guide.
- **Chapter 6:** This chapter concludes this documentation and suggests intended future improvements to be done to maximize the project's potential.

2- Literature Overview

2.1 Project Overview

In recent years, many mobile applications were launched to help people deal with their emotions. These applications typically recommend to the user music or movies or feed-back that can try to improve their emotional state. Existing applications include Behavioral Signals, Morph-Cast, Good Vibrations Company. But these applications are not helping people translate and deliver their emotions to others. In this project, we tried to help people to express their feelings to others in a visual way without directly showing themselves. Recently, various SER techniques employing Signal analysis of sound have been proposed. The advancement of methodologies in SER has been significantly propelled by the emergence of Convolutional Neural Network (CNN) and Transformers, marking a pivotal shift in the field's capabilities. CNN and Transformers improved accuracies of various kinds of Speech recognition tasks such as emotional detection. This chapter discusses some of the approaches employed to solve the tasks of SER. Also, some additional works are discussed relating to the creation of a specific avatar for each detected emotion without the use of visual signals from a camera.

2.2 Theoretical Background

The very first approach for determining the emotional state of a person from his/her speech was made in the late 1970s by Williamson [1], He provided a speech analyzer for the determination of an individual's underlying emotion by analyzing pitch or frequency changes in the speech pattern [2].

Later, in 1996, Dellaert F. *et al.* [3] published the first research paper on the topic and introduced statistical pattern recognition techniques in speech emotion recognition. Dellaert F. *et al.* [3] implemented K-Nearest Neighbors (KNN), Kernel Regression (KR), and Maximum likelihood Bayes' (MLB) classifier using pitch characteristic of the utterances for the recognition of four different emotions (happiness, fear, anger, and sadness). Along with MLB and Nearest Neighbor (NN), Kang B.S. *et al.* [4] implemented the Hidden Markov Model (HMM), where HMM performed the best with 89.1% accuracy for recognizing (happiness, sadness, anger, fear, and neutral) emotions utilizing energy features. Onward, HMM has been largely used by researchers for SER showing satisfactory results. Support Vector Machine (SVM), Gaussian Mixture Model (GMM), and Decision Tree (DT) are some more traditional machine learning models which have been reliably used over the years for the same purpose. In the 2000s, NN has also been widely used for SER studies. Indeed, in the earlier approaches, the use of conventional machine learning algorithms was widespread for recognizing the underlying emotion in human speech [2].

However, in the last decade, the trend of using conventional machine learning models for the recognition of emotion from human speech has moved towards deep learning models. Therefore, deep learning approaches have become more popular, showing promising results. Deep learning algorithms are neural networks with multiple layers. CNN, Deep Convolutional Neural Network (DCNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BLSTM), and Recurrent Neural Network (RNN) are some widely implemented deep learning techniques for SER [2].

In recent years, multitask learning and the attention mechanism are also being used for improved performance and have gained popularity due to their applications to

different learning problems. For cross-corpus and cross-lingual speech emotion recognition, transfer learning techniques are widely used. Another common solution that emerged recently is to combine the multi-head-self-attention-based method with the CNN to train models [2].

Avatar generation refers to the process of creating digital representations of individuals, often used in various applications such as social media, gaming, virtual reality, and online communication platforms. Moreover, these methodologies find extensive application across diverse fields such as Psychology and Anthropology, facilitating a deeper comprehension of human psychology and culture. Leveraging these insights, we can glean valuable understandings into the perception and preferences of avatars, enriching their representation and enhancing their effectiveness in various interactive contexts. Factors such as cultural norms, gender representation, and emotional expression influence how avatars are designed and perceived by users. Machine learning and artificial intelligence techniques are increasingly being used in avatar generation. Generative models such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) can be trained on large datasets of human faces to generate new, realistic faces. These models learn underlying patterns and features of human faces and can generate novel avatars with desired characteristics [5].

2.3 Related Work

This section discusses the commonly used approaches in the Speech Emotion Recognition and Avatar Generation fields along with their methodologies and results as reported by the original papers. The following subsections discuss them in detail as subsection 2.3.1 outlines papers related to the speech emotion recognition task while subsection 2.3.2 discusses papers related to the avatar generation task.

2.3.1 Speech Emotion Recognition

SER is a field focused on identifying and categorizing emotions expressed in spoken language. Using signal processing and machine learning techniques, SER systems analyze audio signals to detect emotional cues. Recent trends in SER involve the adoption of deep learning models like CNNs and Transformers, which have improved accuracy. Additionally, there's a growing interest in applying SER across diverse domains such as human-computer interaction, healthcare, and entertainment. We will proceed to review the most significant recently published papers on the topic.

In [6], Talen Chen *et al.* provided a comprehensive overview and analysis of various machine learning algorithms employed in SER systems. The paper delves into the significance of SER in understanding human emotions expressed through speech and its applications in diverse fields such as healthcare, education, and human-computer interaction. Talen Chen *et al.* meticulously review and compare different machine learning techniques, including SVM, Artificial Neural Networks (ANN), and HMM, highlighting their strengths, weaknesses, and performance in SER tasks.

SVMs are highlighted for their effectiveness in handling high-dimensional data and finding optimal hyperplanes to classify emotions based on extracted features from speech signals, such as Mel-Frequency Cepstral Coefficients (MFCCs) or prosodic features. **ANNs**, particularly deep learning models like CNNs and RNNs, are discussed for their ability to automatically extract hierarchical features from raw speech signals, enabling them to capture complex dependencies in emotional expression. The paper likely compares these algorithms in terms of their strengths such as SVM's robustness with small datasets and ANNs' superior performance

with large datasets—and weaknesses, such as scalability issues with SVMs and computational intensity with ANNs. Moreover, the review explores the application of **HMMs** in modeling temporal dependencies in speech features across time frames, making them suitable for tasks where emotion expression unfolds over time. It discusses HMMs' strengths in segmenting speech into emotional states and their limitations, such as assumptions of feature independence and challenges in capturing long-range dependencies.

The paper evaluates these algorithms' performance in SER tasks using metrics like accuracy, precision, recall, and F1 score, providing insights into how each algorithm performs across different emotional states and datasets. Additionally, it addresses challenges such as variability in emotional expression, cross-cultural differences, and noise in speech signals, while highlighting future directions for research. These include integrating multimodal data sources, enhancing robustness in noisy environments, and improving real-time processing capabilities to advance the state-of-the-art in Speech Emotion Recognition.

Overall, this paper serves as a valuable resource for researchers and practitioners seeking insights into the methodologies and advancements in SER. It offers a critical analysis of SVMs, ANNs, HMMs, and other algorithms, aiming to guide future research efforts towards more accurate and efficient emotion recognition from speech signals in real-world scenarios.

Following the successful application of vision transformers in multiple computer vision tasks, these models have drawn the attention of the signal processing community. This is because signals are often represented as (Discrete Fourier Transform) which can be directly provided as input to vision transformers. However, naively applying transformers to spectrograms is suboptimal as mentioned by Radu Tudor *et al.* in [7]. Since the axes represent distinct dimensions

(frequency and time). Radu Tudor *et al.* argue that a better approach is to separate the attention dedicated to each axis. To this end, the authors propose the Separable Transformer (SepTr), an architecture that employs two transformer blocks in a sequential manner, the first attending to tokens within the same time interval, and the second attending to tokens within the same frequency bin. They conduct experiments on three benchmark datasets (ESC-50, Speech Commands V2, CREMA-D), showing that their separable architecture outperforms conventional vision transformers and other state-of-the-art methods as shown in **Table 2.1**, **Table 2.2**. Unlike standard transformers, SepTr linearly scales the number of trainable parameters with the input size, thus having a lower memory footprint.

Table 2.1: Results of SepTr and its ablated versions in comparison with various state-of-the-art methods on CREMA-D [7].

Method	Accuracy
GRU (Shukla et al. [24])	55.01%
GAN (He et al. [19])	58.71%
ResNet-18 (Georgescu et al. [18])	65.15%
ResNet-18 ensemble (Ristea et al. [22])	68.12%
ViT (Gong et al. [13])	67.81%
SepTr-V	65.29%
SepTr-H	65.11%
SepTr-HV	70.31% [†]
SepTr-VH (proposed)	70.47%[†]

Table 2.2: Results of SepTr versus various state-of-the-art methods on Speech Commands V2 (SCV2) and ESC-50 [7].

Method	SCV2	ESC-50
RBM (Sailor et al. [23])	-	86.50%
EfficientNet (Kim et al. [20])	-	89.50%
MatchboxNet (Majumdar et al. [21])	97.40%	-
ViT (Gong et al. [13])	98.11%	88.70%
SepTr (proposed)	98.51%[†]	91.13%[†]

As conclusion from [Table 2.2](#), it is clear that the highest accuracy shown is using SepTr using dataset SCV2.

In [8], Fakhar Anjam *et al.* focus on one of the key challenges in speech emotion recognition which is the extraction of the emotional features effectively from a speech utterance. Despite the promising results of recent studies, they generally do not leverage advanced fusion algorithms for the generation of effective representations of emotional features in speech utterances. To address this problem, as mentioned in the paper they describe the fusion of spatial and temporal feature representations of speech emotion by parallelizing CNNs and a Transformer encoder for SER. Fakhar Anjam *et al.* stack two parallel CNNs for spatial feature representation in parallel to a Transformer encoder for temporal feature representation, thereby simultaneously expanding the filter depth and reducing the feature map with an expressive hierarchical feature representation at a lower computational cost as shown in [Figure 2.1](#).

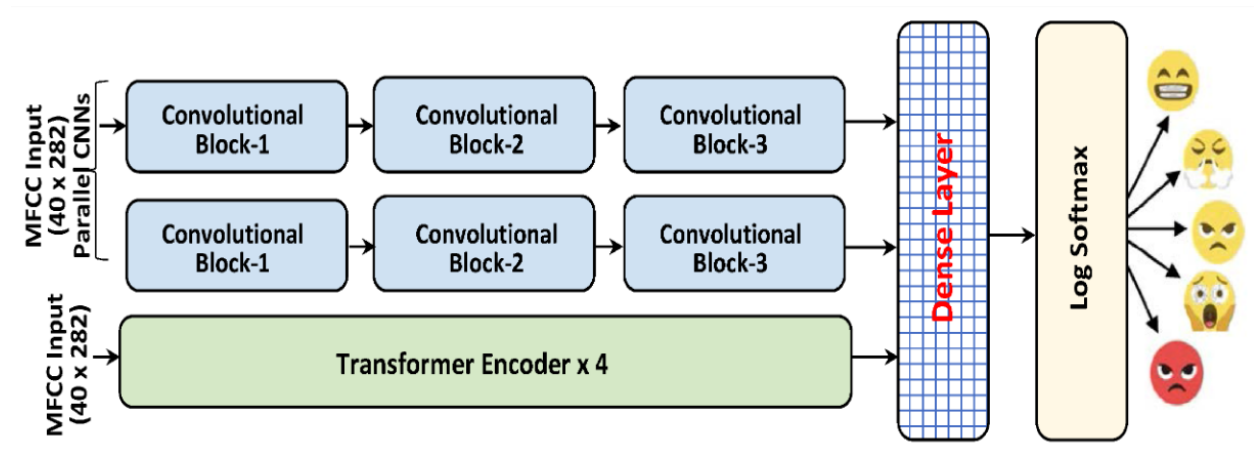


Figure 2.1: Two parallel CNNs with Transformer encoder for feature extraction. The extracted features are fed to the dense layer with a log SoftMax classifier for emotional state prediction.

Yassin Ben Ayed *et al.* [9] delve into fundamental aspects of SER such as data representation, where raw audio signals are transformed into spectrograms or

MFCCs. Spectrograms capture the frequency content over time, while MFCCs represent the short-term power spectrum, both pivotal in enabling neural networks to discern emotional nuances from speech.

In terms of model architectures, the paper may explore various approaches like CNNs, RNNs such as LSTM, or hybrid models like CNN-RNNs. These architectures are meticulously designed to capture temporal dependencies and hierarchical features within speech data, optimizing the recognition of emotional patterns.

Training strategies are crucially discussed, encompassing the utilization of annotated datasets to optimize model parameters via algorithms like stochastic gradient descent. Evaluation metrics such as accuracy, precision, recall, and F1-score gauge the efficacy of these models in accurately identifying emotions, reflecting their robustness and reliability in practical applications.

The paper addresses inherent challenges in SER, including variability in emotional expression, environmental noise, and limited labeled data. Proposed solutions often include advanced techniques like data augmentation to enrich datasets, transfer learning from pre-trained models to leverage existing knowledge, or innovative network architectures tailored to enhance performance under real-world conditions. Ultimately, the paper explores the wide-ranging applications of SER across diverse domains, from emotion-aware virtual assistants and sentiment analysis in customer service to emotion monitoring in healthcare settings. By deciphering emotional cues from speech, SER not only enhances human-computer interaction but also augments user experience by fostering more intuitive and responsive technologies.

In conclusion, the paper underscores the transformative potential of deep learning in advancing Speech Emotion Recognition, offering insights into cutting-edge methodologies and their applications. By leveraging neural networks to decode

emotional states from speech, researchers aim to enrich the functionality and impact of SER systems across various societal and commercial applications.

Finally, **Table 2.3** shows a summarized overview of each paper discussed in this section that highlights the model, datasets and results obtained in each paper. From this comparison, we found that the model that achieved the highest accuracy was separable transformer [7]. We also observed that when [6] and [8] used the same dataset (RAVADESS), [8] achieved the higher accuracy.

Table 2.3: Summarized Overview of Significant SER Papers

Paper	Model	Dataset	Evaluation (Accuracy)
Separable Transformer [7]	Separable Transformers	- CERMA-D - ESC-50 - Speech Commands	70.47% 91.13% 98.51%
SER using machine learning [6]	CNN	- RAVADESS	80%
SER using CNN and multi head convolutional transformer [8]	Convolution Neural Networks and Multi- Head Convolution Transformer	- RAVDESS - IEMOCAP	82.31% 79.42%
SER using Deep Learning [9]	SVM and Auto Encoder	- RML	65.43%

2.3.2 Avatar Generation

There are different techniques for avatar generation such as: Image-to-Image Translation techniques, matching realistic photograph components to cartoon-like avatar components technique, and the Text-and-Shape guided 3D Human Avatar Generation via Diffusion Models technique. We focused in our survey on Image-to-Image Translation as it is usually used for changing the expression or parts in the photo because we wanted to create different images with different emotions to visualize the detected emotion in the voice. Most of papers use Generative Adversarial Networks (GANs) in Unsupervised Image-to-Image Translation from different domains with unpaired image data. But most of them create images with low-quality and some of them change unwanted parts of the input image which cause unwanted image as an output. The main goal is to create realistic-looking avatars that resemble a person with different emotions. It works by analyzing various facial features, such as the shape of the face, eyes, nose, and mouth, to know the most important parts to change and then generating avatars with different emotions.

In [10], Tang *et al.* propose a solution that takes photo of human face as an input and generates seven photos as an output that show seven different emotions (angry, contemptuous, disgust, fearful, sad, happy, and surprised). The authors explained that the use of GANs sometimes generated images with low-quality, so they proposed an architecture that focuses on the generation of images with high-quality. GANs are able to learn a mapping function from one image domain to another with unpaired image data but they are only able to convert low-level information but fail to transfer high-level semantic parts of images as they do not have the ability to detect the most discriminative semantic parts of images, which thus generates low-quality images. Alternatively, Attention-Guided Generative

Adversarial Networks (AGGAN) can detect the most discriminative semantic object and minimize changes of unwanted part for semantic manipulation problems without using extra data and models.

The attention-guided generators in AGGAN can produce attention masks via a built-in attention mechanism, and then fuse the input image with the attention mask to obtain a target image with high-quality. Tang *et al.* proposed a novel attention-guided discriminator which only considers attended regions. And AGGAN is trained in an end-to-end fashion with adversarial loss, cycle-consistency loss, pixel loss and attention loss that make them more effective to generate sharper and more accurate images than existing models. The attention-guided generators focus only on those regions of the image that are responsible for generating the novel expression such as eyes and mouth and keep the rest parts of the image such as hair, glasses, and clothes untouched. The higher intensity in the attention mask means the larger contribution for changing the expression as shown in **Figure 2.2**. **Table 2.4** reports the results on AR Face, Bu3dfe and CelebA datasets which show that the proposed AGGAN achieves the best results on these datasets compared with competing models.

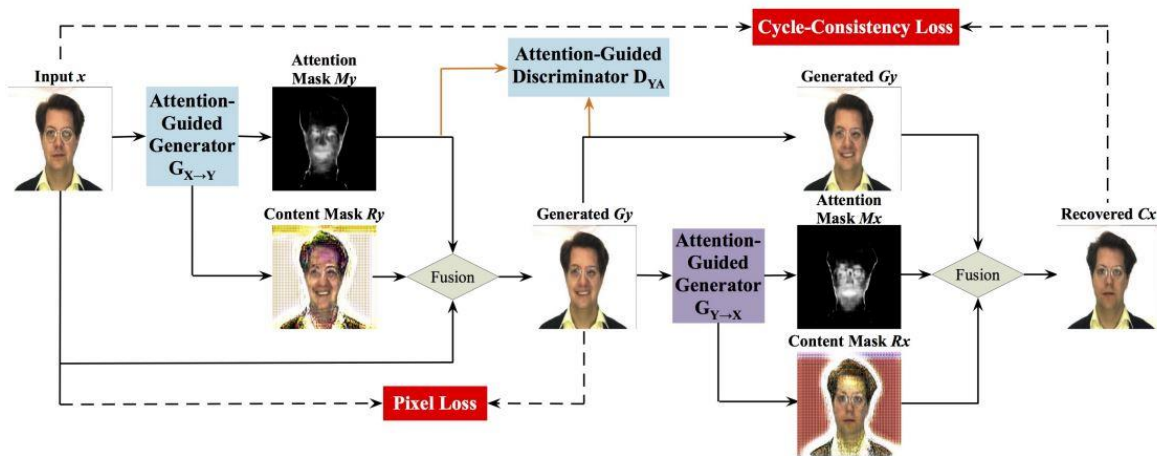


Figure 2.2: The attention-guided generators process [10]

Table 2.4: Quantitative comparison with different models. For all metrics except MSE, higher is better2.4 [10]

Model	AR Face			Bu3dfe			CelebA
	AMT	PSNR	MSE	AMT	PSNR	MSE	AMT
CycleGAN	10.2	14.8142	2538.4	25.4	21.1369	602.1	34.6
DualGAN	1.3	14.7458	2545.7	4.1	21.0617	595.1	3.2
DiscoGAN	0.1	13.1547	3321.9	0.2	15.4010	2018.7	1.2
ComboGAN	1.5	14.7465	2550.6	28.7	20.7377	664.4	9.6
DistanceGAN	0.3	11.4983	4918.5	8.9	13.9514	3426.6	1.9
Dist. + Cycle	0.1	3.8632	27516.2	0.1	10.8042	6066.7	1.3
Self Dist.	0.1	3.8674	26775.4	0.1	6.6458	14184.2	1.2
StarGAN	1.6	13.5757	3360.2	5.3	20.8275	634.4	14.8
ContrastGAN	8.3	14.8495	2511.1	26.2	21.1205	607.8	25.1
Pix2pix	2.6	14.6118	2601.3	3.8	21.2864	580.6	-
Enc.-Decoder	0.1	12.6660	3755.4	0.2	16.5609	1576.7	-
BicycleGAN	1.5	14.7914	2541.8	3.2	19.1703	1045.4	-
AGGAN	12.8	14.9187	2508.6	32.9	21.3247	574.5	38.9

On the other hand, there are tools that generate customized avatars based on choices for the shape of the eye, nose, mouse, brows, hair, and clothes. You can add a beard, or glasses. It faster than other techniques but provides limited options when compared with deep learning generation-based methods.

3- System Architecture and Methods

3.1 System Architecture

Figure 3.1 shows our system architecture (3-Tier). It consists of 3 layers; presentation layer, logic layer, and data layer. The presentation layer contains the application interface where the user enters a recorded audio then it is passed to the logic layer. In the logic layer, the audio goes through preprocessing module where MFCC features are extracted from it and segmentation is applied on it to segment the long audio into smaller statements then we use the trained SER module saved in the data layer to detect emotions of the segmented statements. The detected emotions are passed to avatar generation module where the avatar of the detected emotions is generated and finally the video is generated.

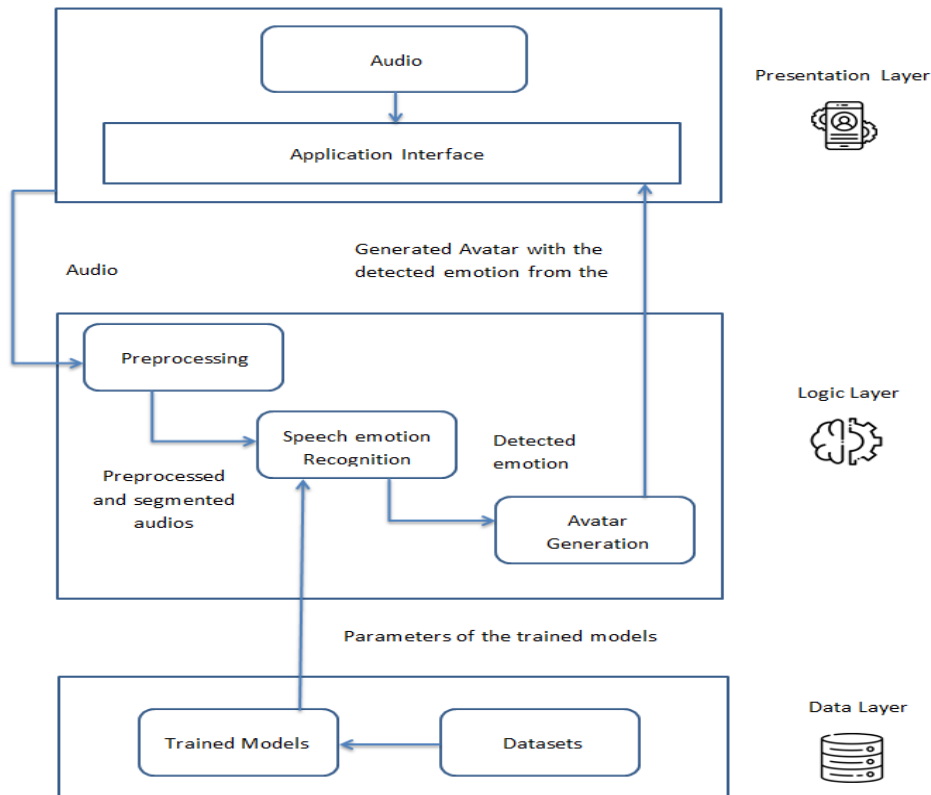


Figure 3.1: System Architecture

3.2 Description of Methods and Procedures

3.2.1 Presentation Layer

User Interface and communication layer of the application where end-user interacts with the application. In the Application Interface, a recorded audio is taken from the user, or the user can record a new audio in the application. This audio is then sent to the logic layer to be processed.

3.2.2 Logic Layer

The audio is preprocessed before sending it to the speech emotion recognition module. We split audio into small audio segments according to silence. Then remove noise from audio. Then detect emotions from audios according to MFCC. The detected emotions are then sent to the Avatar Generation module to create an emotional avatar. Lastly, this avatar is sent to the presentation layer for the user to see the video.

3.2.2.1 Speech emotion recognition

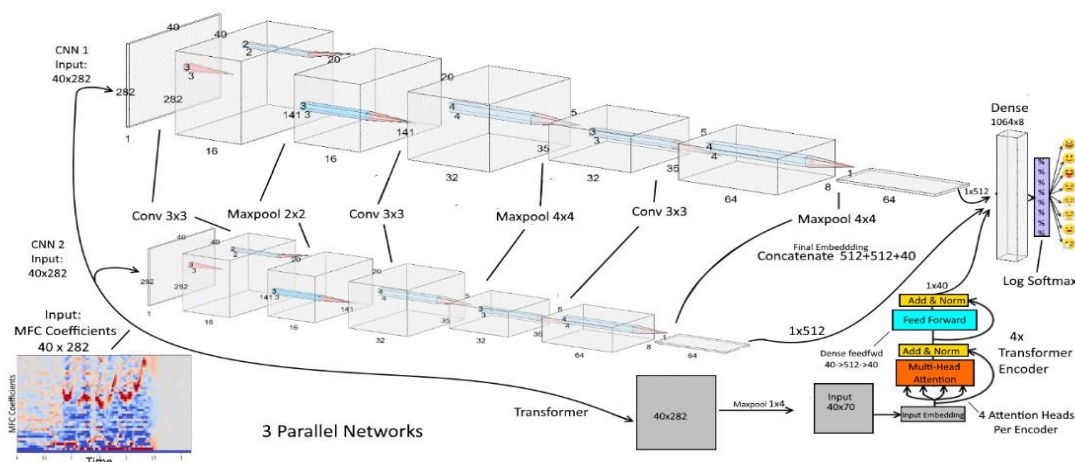


Figure 3.2: Panoramic view of the SER model used.

Figure 3.2 shows a panoramic view of the final SER model used in our application. It follows [11] as it contains two parallel CNN model and multi-head attention transformer encoder to recognize emotions in speech spectrograms from audio.

Figure 3.3 shows the details of the two parallel CNN networks used within the SER model in Figure 3.2. Each CNN model consists of a 3-layer deep 2D convolutional block. Similar to the classic LeNet architecture. This models' CNN consists of the following order of layers: convolutional block → max pooling → convolutional block → max pooling → fully connected.

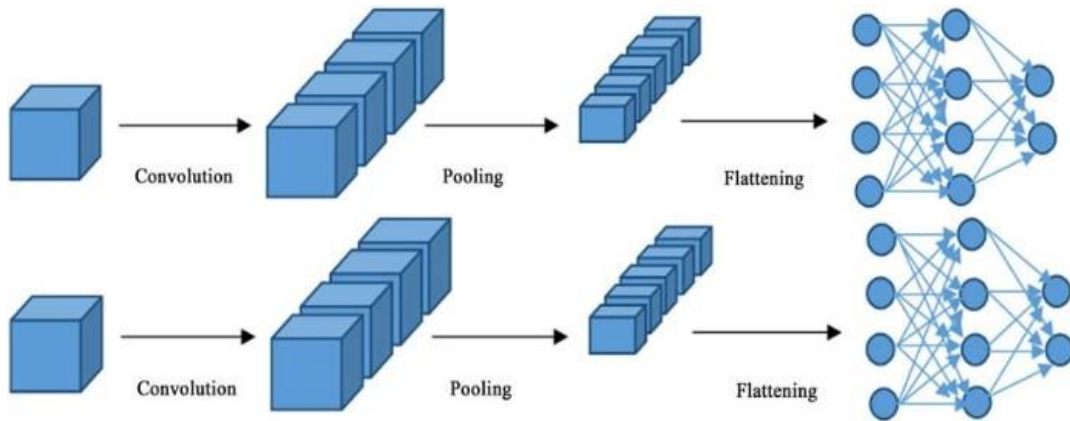


Figure 3.3: Two parallel CNN

Transformer-Encoder

Figure 3.4 shows transformer encoder used within the SER model in figure 3.2.

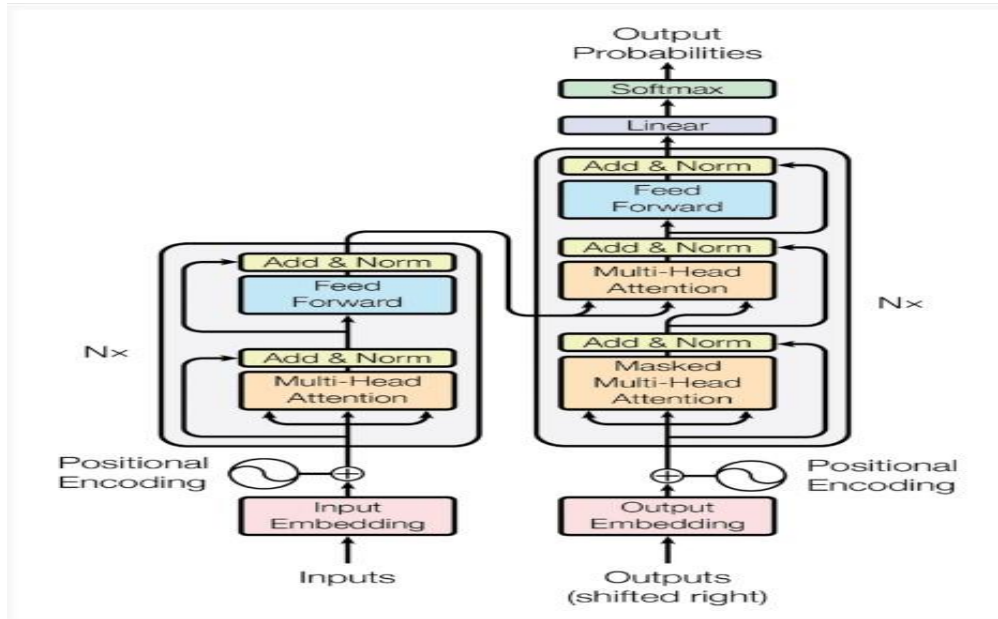


Figure 3.4: Transformer-Encoder.

When MFCCs enter a transformer encoder, they are first extracted from the raw audio signal, representing the audio's essential characteristics. These MFCCs, organized as a matrix of time frames and coefficients, may be segmented into fixed-size chunks. Each segment is fed into the transformer encoder, where positional encoding is added to provide temporal context. The transformer encoder, consisting of layers of self-attention and feed-forward neural networks, processes these feature vectors. The self-attention mechanism helps the model capture patterns and dependencies within the audio sequence. The encoder refines the feature vectors into rich, abstract representations, which can be used for various audio processing tasks.

3.2.2.2 Avatar Generation

We generated the appropriate avatars by taking the emotions detected from the SER module and creating the suitable combinations of attributes such as : raised

eyebrows , teary eyes , downturned mouthetc., to reflect the suitable emotion detected from the SER module .Then we sent these combinations to python-avatar library to generate an appropriate avatar. There are many available attributes such as:

- style
- background_color
- eyes
- top
- eyebrows
- nose
- facial_hair
- hair_color
- accessory
- clothing
- clothing_color
- mouth

To create the emotional avatar, we focused on the attributes (mouth, eyes, eyebrows).

We also created 2 copies of this emotional avatar:

- One with closed mouth
- One with open mouth

So that it would give the sense of speaking .To create the video, all emotions detected from the audio is put in an array and displayed along with the video.

3.2.3 Data Layer

The data layer in a system architecture acts as a robust infrastructure layer that manages the lifecycle of data, from ingestion to storage to consumption, while ensuring reliability, performance, and compliance with regulatory requirements. It contains a dataset of audios.

3.3 System Users

An efficient model that detects emotions from voice and visualizes these emotions through an avatar does not require extra technical expertise and can be used in many fields:

Medical field: can help people who have social phobias to communicate easily with others.

Games: Deliver a virtual reflection of a player's emotions while protecting their privacy.

Customer service: Can aid the call center workers.

E-learning and virtual meetings: Help the teachers to detect the state of the students throughout the sessions.

Supports privacy: You can use a photo that isn't yours or by using a default avatar.

4- System Implementation and Results

4.1 Datasets

Three datasets are used to train and evaluate different models in our project with the purpose of detecting emotions from speech. Each dataset is discussed below, the datasets are: RAVDESS [12], SAVEE [13] and TESS [14].

RAVDESS [12]:

The dataset "RAVDESS" contains 1440 files of emotional speech audio (60 trials per actor x 24 actors). It contains 24 professional actors (12 female, 12 male). Speech emotions include calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics.

File name identifier.

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised)
- Emotional intensity (01 = normal, 02 = strong).
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

CNN and Multi-Head Convolutional Transformer [8] model is trained on this dataset to detect emotion from speech and as a result the dataset has been separated into 1152 training, 144 validation and 144 testing.

SAVEE [13]:

The dataset "SAVEE" contains 480 files of emotional speech audio including 384 training, 48 validation and 48 testing. It is male only and very high-quality audio. It was recorded by four native English male speakers. Emotion has been described psychologically in discrete categories: (anger, disgust, fear, happiness, sadness, and surprise). A neutral category is also added to provide recordings of 7 emotion categories. The filename consists of 3 identifiers (e.g., DC_d04.wav).

Filename identifier :

- English male speakers (DC, JE, JK, KL).
- Emotion (the first character from 7 emotions).
- Number of sentence (The text material consisted of 15 TIMIT sentences per emotion).

TESS [14]:

The dataset "TESS" contains 2800 files of emotional speech audio including 2240 training, 280 validation and 280 testing. It is female only and is of very high-quality audio. There are a set of 200 target words that were spoken in the carrier phrase. And recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). The filename consists of 3 identifiers (e.g., YAF_back_happy.wav).

Filename identifier:

- Name of character.
- Main word
- Emotion

Table 4.1: Used datasets comparison.

Name	Speakers	Genders	Size	Emotion
RAVDESS [12]	24 professional actors (12 female, 12 male)	Both(female and male)	1440 files	8 emotions
SAVEE [13]	four native English male speakers	Male only	480 files	7 emotions
TESS [14]	two actresses (aged 26 and 64 years)	Female only	2800 files	7 emotions

Table 4.1 shows a comparison between the datasets we used in our experiments throughout the project, it is clear from this comparison that RAVDESS [12] has the advantage of containing both genders (males and females) and 8 emotions. It is also a very clean dataset compared to the other datasets but on the other hand it is the smallest one, so we used augmentation before preprocessing step. SAVEE [13] dataset is male only and is of very high-quality audio, on the other hand TESS [14] dataset is female only and is of very high-quality audio. Most of the other dataset is skewed towards male speakers and thus brings about a slightly imbalance representation. So, because of that, this dataset would serve as a very good training dataset for the emotion classifier in terms of generalization.

4.2 Software Tools Used

Python: Python is a programming language that lets you work quickly and integrate systems more effectively .In the project, certain frameworks and applications of Python are leveraged, such as:

- **PyTorch:** Used in building recognition model (CNN and Multi-Head Convolutional Transformer).
- **Flask:** Used in building REST API for the application.
- **PyDub:** Used in segmentation of audio into chunks.
- **Python-avatar:** Used to generate avatars.
- **Pyaudio :** Used in realtime for audio recording.

Android: Android is a mobile operating system based on a modified version of the Linux kernel and other open-source software, designed primarily for touchscreen mobile devices such as smartphones and tablets.

Google colaboratory: Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs.

4.3 Setup Configuration (software)

Flask: Flask is a popular web framework in Python used for building web applications. It provides a simple and flexible way to handle HTTP requests, define routes, and render templates.

Pandas: Pandas is a powerful library for data manipulation and analysis in Python. It provides data structures and functions for efficiently handling and analyzing structured data, particularly tabular data.

NumPy: NumPy is a fundamental package for scientific computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a wide range of mathematical functions to operate on these arrays efficiently.

PyDub: PyDub is a python library that provides a gold mine of tools for manipulating audio files. It becomes a programmatic way to ensure the audio files are consistent and in an ideal format for transcription locally or through an API.

Python-avatar: Python-avatar it is a Python package that allows the generation of avatar images programmatically. These avatars are often used in applications, websites, or any other place where you need to represent users or entities with visual icons. The library typically provides various options for customization, such as different styles, colors, and features.

CairoSVG: CairoSVG it is a Python library for converting SVG images to other formats like PNG, PDF, SVG, or PostScript using the Cairo graphics library. It allows the work with SVG files in Python, which is particularly useful when dealing with scalable graphics in web development, data visualization, or any application that requires vector images.

Pyaudio: Pyaudio it is a Python library that provides Python bindings for PortAudio, the cross-platform audio I/O library. With PyAudio, you can easily use Python to record audio from a microphone, play and record audio on a variety of platforms and real-time audio processing. PyAudio requires PortAudio, which may need to be installed separately depending on your operating system. **Precompiled binaries are available for Python 3.7, 3.8, and 3.9. For newer versions, you might need to compile from source or use unofficial binaries.**

4.4 Experiments and Results

This section discusses the experiments performed in the project and the results obtained. These experiments were mostly performed in the preprocessing and SER module.

4.4.1 Speech emotion recognition experiments

Two models were applied to detect emotions based on the tone of the speaker. Each model is explained below along with the experiments and results obtained for each model. These two models are LSTM [15] and, CNN and Multi-Head Convolutional Transformer [8].

1- LSTM model [15]:

This model contains multi layers from LSTM (Long Short-Term Memory) neural network architecture to recognize emotions in audio data. The model is based on an original paper [15] and employs a combination of three datasets:

- RAVDESS [12]
- TESS [14]
- SAVEE [13]

Data Preprocessing and Augmentation

Before training, the data undergoes extensive preprocessing, including:

- Loading and Preprocessing: The three datasets are combined, and audio files are processed for consistency.
- Visualization: Wave plots and spectrograms are generated to visualize different emotions.

Data Augmentation: Techniques such as noise addition, time stretching, and pitch shifting are applied to increase dataset diversity.

➤ **Noise Addition:**

- **Description:** Adds random noise to the audio to simulate real-world conditions.
- **Types:** White noise, Gaussian noise, environmental noise.

➤ **Time Stretching:**

- **Description:** Changes the speed of the audio without altering its pitch.
- **Types:** Uniform and non-uniform stretching.

➤ **Pitch Shifting:**

- **Description:** Alters the pitch of the audio without changing its duration.
- **Types:** Uniform and non-uniform shifting.

Model Performance

The model achieved an accuracy of 71%, while the original paper [15] achieved an accuracy of 72%.

Figure 4.1 shows the training and validation loss curves. We used this visualization to validate that the model was not overfitting and for early stopping.

Figure 4.2 shows the confusion matrix obtained by this model. We observed that the neutral and unpleasant emotions were the best emotions the model could predict.



Figure 4.1: Loss curve for LSTM model [15]

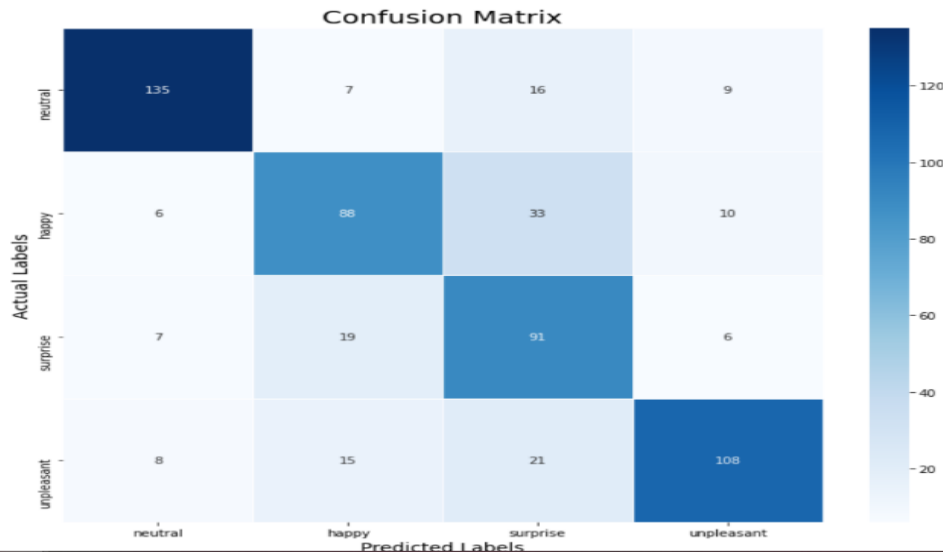


Figure 4.2: Confusion matrix for LSTM model [15]

In addition, several trials have been applied to reach the best version of the model. These trials were related to Preprocessing, architecture, and dataset. Each experiment is explained below:

Preprocessing:

The model is based on the original paper [15]. We used 8 emotion classes instead of 4. The original paper dropped rows where emotions were fear or disgust because the dataset (SAVEE) did not contain these emotions, also dropped rows where emotions were sad or angry and replaced them with unpleasant. Instead of 4 emotions {neutral, happy, surprise, unpleasant}. We used 8 emotions { neutral, happy, sad, angry, disgust, surprise, calm, fearful }.

This experiment model achieved an accuracy of 62%.

Architecture:

- First, we added a convolution layer followed by Batch Normalization layer. Convolutional layers are powerful for extracting spatial features from data, particularly useful for capturing local dependencies and patterns. By adding a convolutional layer, we aimed to enhance the model's ability to process and understand the input data at a finer granularity. Batch normalization was added after the convolutional layer to standardize the inputs to the next layer, speeding up training and improving model stability. This addition helped in capturing more intricate patterns in the data, leading to improved feature representation.
- The batch normalization layer contributed to faster convergence and reduced overfitting, resulting in a more robust model. Secondly, we increased the number of LSTM layers from 2 to 3. LSTM layers are effective at capturing temporal dependencies in sequential data. By increasing the number of LSTM layers, we aimed to allow the model to learn more complex and hierarchical temporal features. This addition helped in capturing more

intricate patterns in the data, leading to improved feature representation. The batch normalization layer contributed to faster convergence and reduced overfitting, resulting in a more robust model, the accuracy of which was 89%. [Table 4.2](#) shows a summarized overview of the trials applied on the LSTM model and the results obtained.

Table 4.2: Architecture changes in LSTM model [15]

Model Changes	Accuracy
Added a convolutional layer Followed by Batch Normalization layer.	76%
Increased number of LSTM layers from 2 to 3.	89%

2- CNN and Multi-Head Convolutional Transformer [8]:

This model is the main model used in this project with RAVDESS [12] dataset because transformer proved to be better than LSTM as it achieved an accuracy of 95%. [Figure](#) shows the training and validation loss. We used this visualization to make sure that the model was not overfitting and for early stopping, [Figure 4.4](#) shows the confusion matrix obtained by this model and it shows that the model couldn't differentiate between the calm and the natural emotions, but the angry and sad emotions are the best emotions the model could predict.

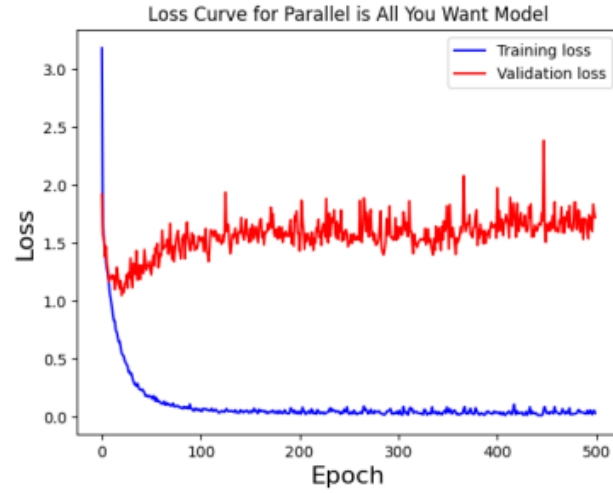


Figure 4.3: Loss Curve for CNN and Multi-Head Convolutional Transformer model [8].

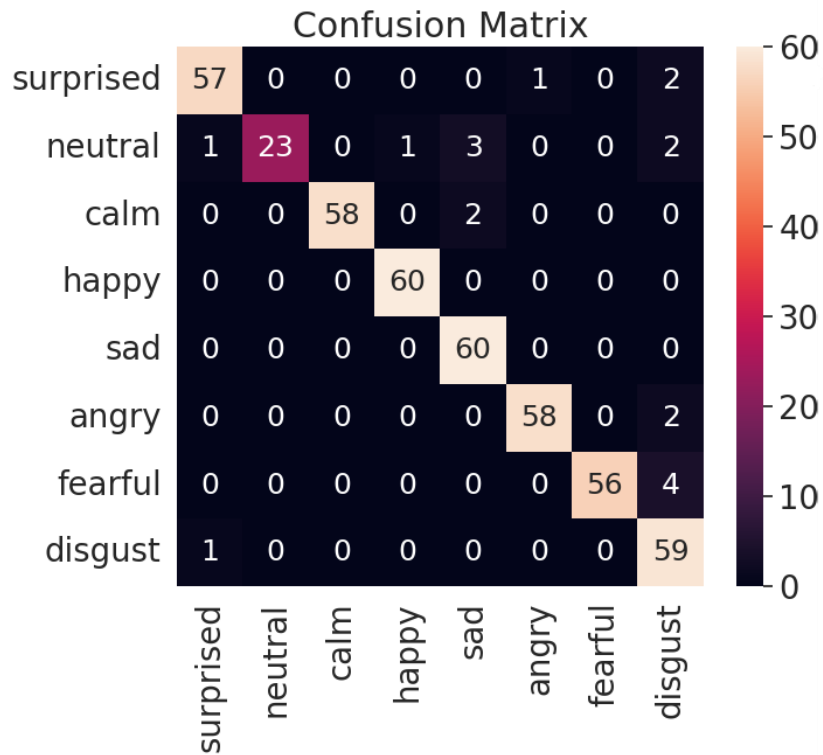


Figure 4.4: Confusion matrix for CNN and Multi-Head Convolutional Transformer model [8] on RAVDESS [12] dataset.

In addition, several trials have been applied to reach the best version of the model. These trials were related to Preprocessing, architecture, and dataset. Each experiment is explained below:

Preprocessing:

In feature extraction, MFCCs and Mel-spectrogram have been experimented with and the results are reported in [Table 4.3](#). Mel-Spectrogram is computed by applying a Fourier transform to analyze the frequency content of a signal and to convert it to the Mel-scale, while MFCCs are calculated with a discrete cosine transform (DCT) into a Mel frequency spectrogram and it is a bit more decorrelated, which can be beneficial with linear models like Gaussian Mixture Models. Shortly, MFCC is more relevant than Mel-spectrogram as evident below.

Table4.3: Preprocessing changes in CNN and Multi-Head Convolutional Transformer model [8].

MFCC	Mel-spectrogram
Accuracy (95%)	Accuracy (75%)

Architecture:

The main architecture contains two parallel CNNs to detect spatial features and a multi-head attention Transformer encoder to detect temporal features. The changes that were implemented in architecture are shown in [Table 4.4](#) .

Table 4.4: Architecture changes in CNN and Multi-Head Convolutional Transformer model [8].

Changes	Accuracy
In transformer encoder block: the number of layers has been changed from 4 to 6.	76%
In CNN: Changed Max_pooling to avarege_pooling.	74%
Used transformer encoder only	78%
Used normalization layer instead of batch normalization	82%

Changes in datasets:

The main dataset used in this model is RAVDESS [12] and other datasets have been experimented with. These datasets are SAVEE [13] and combination from RAVDESS [12] and SAVEE [13]. The results of these experiments are reported below in Table 4.5.

Table 4.5: Datasets trials in CNN and Multi-Head Convolutional Transformer model [8].

Dataset	RAVDESS [12]	SAVEE [13]	RAVDESS [12] + SAVEE [13]
Accuracy	95%	62.5%	72.5%

As shown in Figure 4.5, confusion matrix illustrates that the model with using SAVEE [13] dataset couldn't differentiate between the anger and the disgust emotions and the same between the fear and surprise emotions. As a result, the

model in the light of usage RAVDESS [12] is better than the model in the light of usage SAVEE [13].

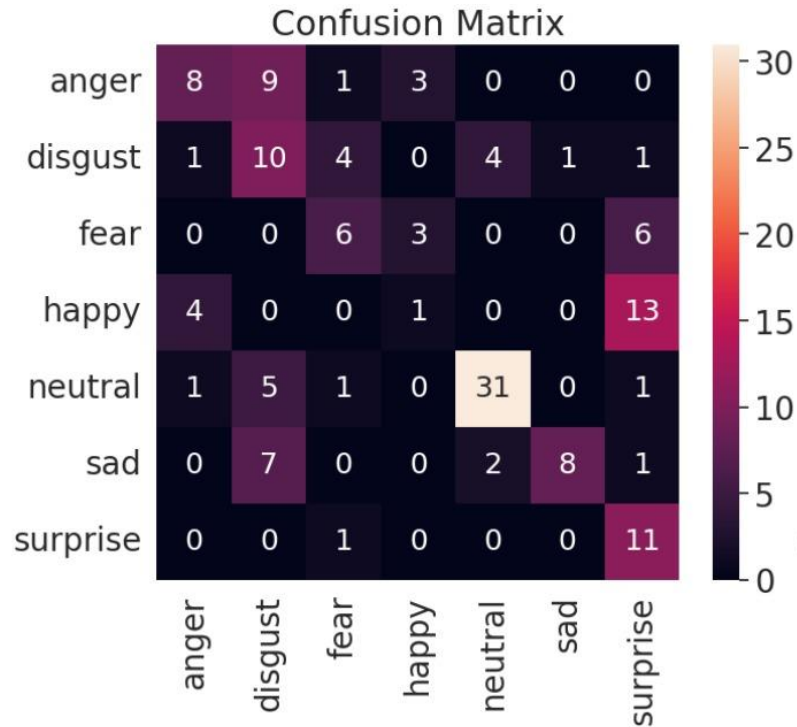


Figure 4.5: Confusion matrix for CNN and Multi-Head Convolutional Transformer model [8] on SAVEE [13] dataset.

As shown in Figure 4.6, confusion matrix results illustrate that combination between RAVDESS [12] and SAVEE [13] datasets in the model are better in the light usage SAVEE [13]. Although the model fails to detect happy and sad emotions, its effectiveness shows up in detecting almost all the rest of the emotion in a correct way.

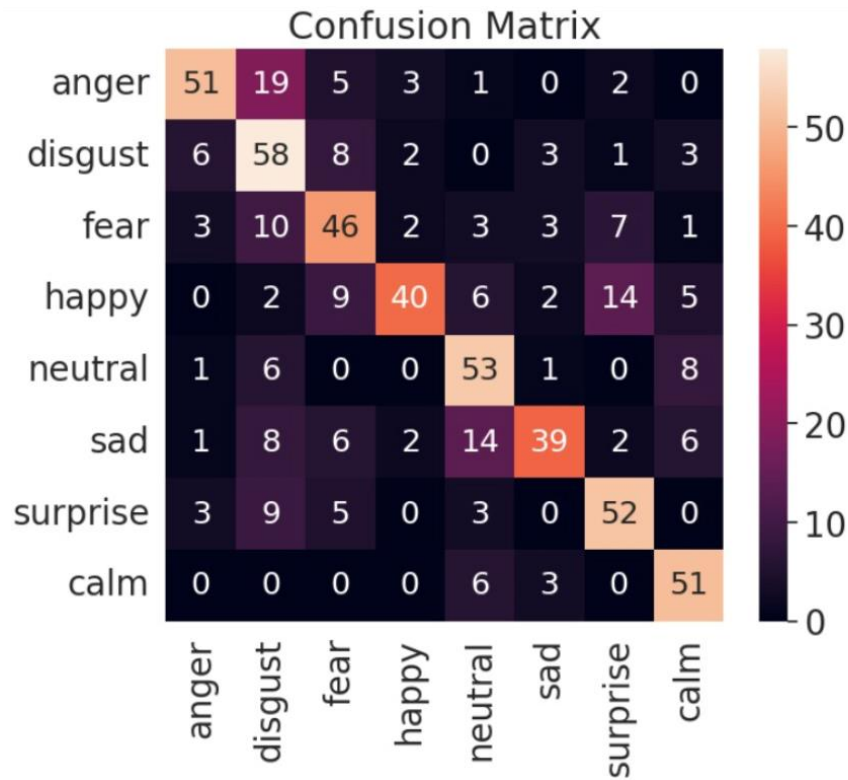


Figure 4.6: Confusion matrix for CNN and Multi-Head Convolutional Transformer model [8] on RAVDESS [12] + SAVEE [13] dataset.

4.4.2 Avatar Generation experiments

We created avatars reflecting emotions detected by the SER module, using attributes like raised eyebrows, eyes, and mouths. The combinations were processed with the python-avatar library. We focused on the **mouth, eyes, and eyebrows** to express emotions.

The configurations applied to simulate the 8 different emotions supported in our application were:

- 1- Calm: eyebrows→default natural, eyes→closed, mouth→twinkle (see figure 4.7).
- 2- Happy: eyebrows→default natural, eyes→happy, mouth→smile (see figure 4.8).
- 3- Disgust: eyebrows→default natural, eyes→side, mouth→concerned (see figure 4.9).
- 4- Sad: eyebrows→sad_concerned_natural, eyes→cry, mouth→sad (see figure 4.10).
- 5- Fearful: eyebrows→default natural, eyes→squint, mouth→ scream open (see figure 4.11).
- 6- Surprise: eyebrows→default natural, eyes→surprise, mouth→disbelief (see figure 4.12).
- 7- Angry: eyebrows→angry, eyes→squint, mouth→grimace (see figure 4.13).



Figure 4.7: Calm

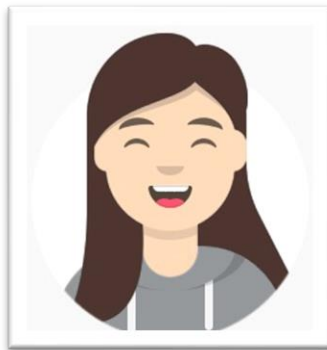


Figure 4.8: Happy

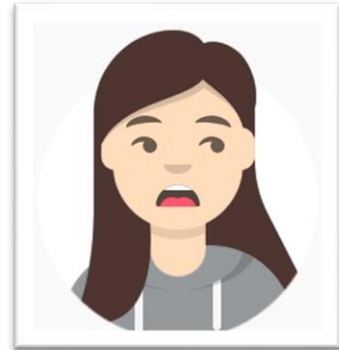


Figure 4.9: Disgust

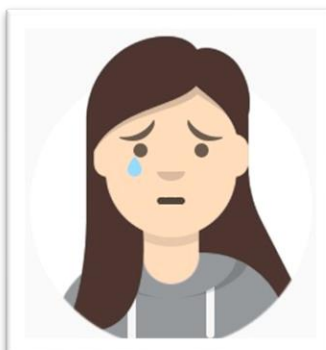


Figure 4.10: Sad

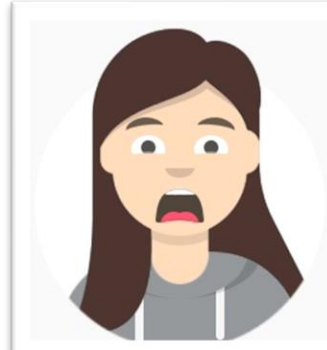


Figure 4.11: Fearful

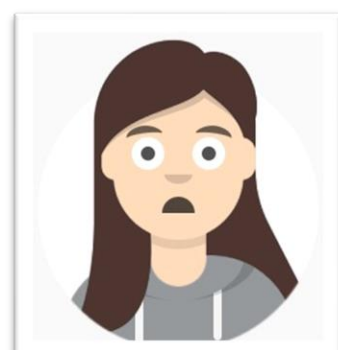


Figure 4.12: Surprise



Figure 4.13: Angry

Each avatar had two versions one with a closed mouth and one with an open mouth to simulate speaking as shown in Figures (4.14) and (4.15). Emotions detected from the audio are compiled into an array and the appropriate avatars are generated and displayed one after the other in a video simulating a speaking person.

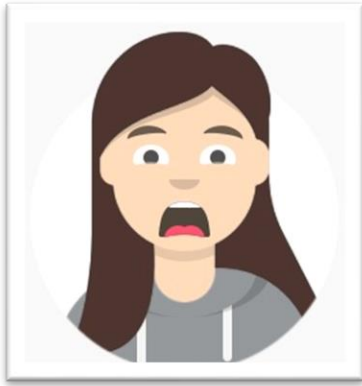


Figure 4.14: Fear with opened mouth.

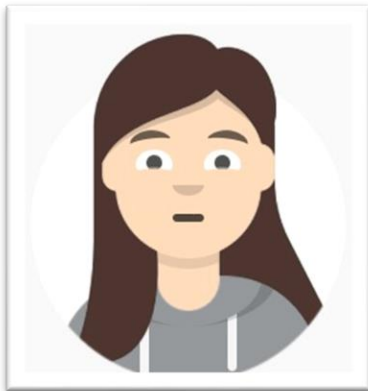


Figure 4.15: Fear with closed mouth.

4.4.3 Realtime

In Realtime, Pyaudio library is used to record audio from a microphone. The recorded audio is saved to the device, then it is passed to our trained model so a small video is generated and this video is displayed. This process is repeated every 10 seconds. You can stop recording through keyboard interrupt. So every 10 seconds the user is able to see a new emotional video to the words you have said 10 seconds later

5- User Manual

This chapter is a walkthrough of the whole application.

To use the application, you have to allow permissions such as internet, read external storage and write to external storage.

* Home page:

- This is the first page in the application which shows two buttons: Login and Register.
- Options are available for users to either log in with existing credentials or register a new account as shown in [Figure 5.1](#).

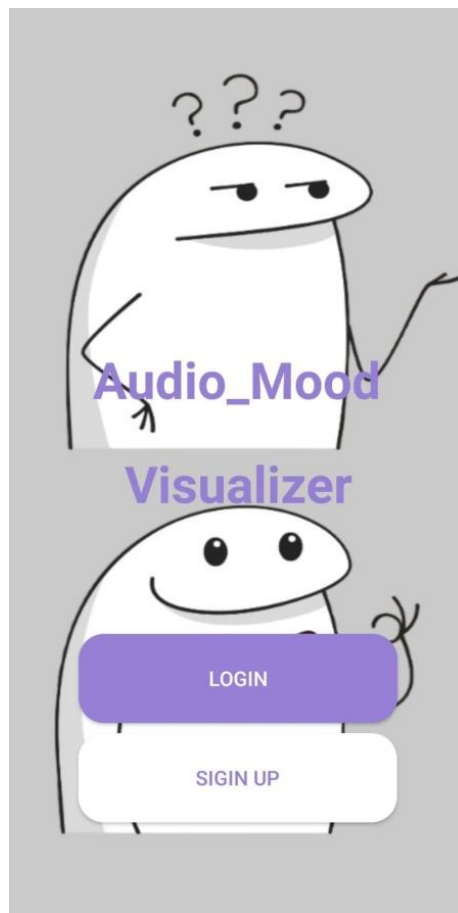


Figure 5.1: Home page

* User Registration:

The registration form includes the following fields:

- Username: A unique identifier chosen by the user.
- Password: A secret word or phrase used by the user to log in.
- Gender: The gender of the user. (Male, Female).
- Phone Number: The user's contact number and must be a valid phone number format.
- Age: The user's age and must be a valid integer.
- Sign up button: check username is unique and strong password show in image of [Figure 5.2](#).

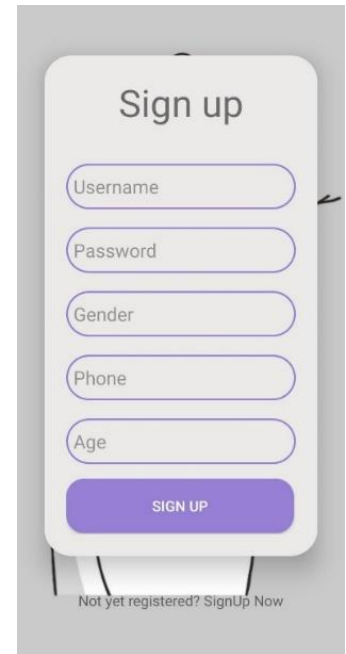
A mobile app interface for user registration. The form is titled "Sign up" and contains five input fields: "Username", "Password", "Gender", "Phone", and "Age". Below the fields is a purple button labeled "SIGN UP". At the bottom, there is a link that says "Not yet registered? Sign Up Now".

Figure 5.2: User Registration

* User Login:

The user form includes the following fields:

- Username: The field where users enter their username.
- Password: The field where users enter their Password.
- Login Button: The button that users click to submit their username and password. Must be correct username and password to login to application. Show in image of [Figure 5.3](#).

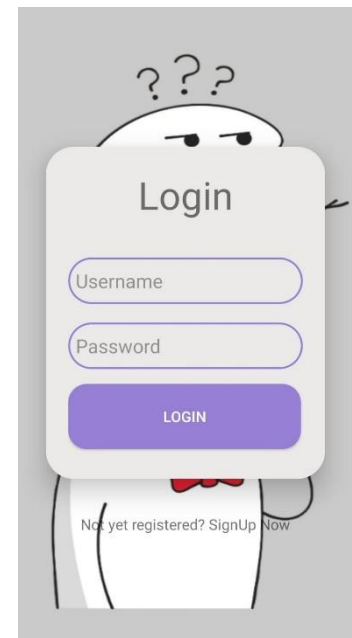
A mobile app interface for user login. The form is titled "Login" and contains two input fields: "Username" and "Password". Below the fields is a purple button labeled "LOGIN". At the bottom, there is a link that says "Not yet registered? Sign Up Now". The form is overlaid on a cartoon character with question marks above its head.

Figure 5.3: User Login

* **Choose avatar:**

- The Avatar Selection page allows users to choose an avatar based on the user type they have previously entered. There are three avatars available for selection, and the chosen avatar will be used to represent the user in a video that corresponds to their voice input. Avatar options for female users are shown in image [Figure 5.4](#) and for male users [Figure 5.5](#).

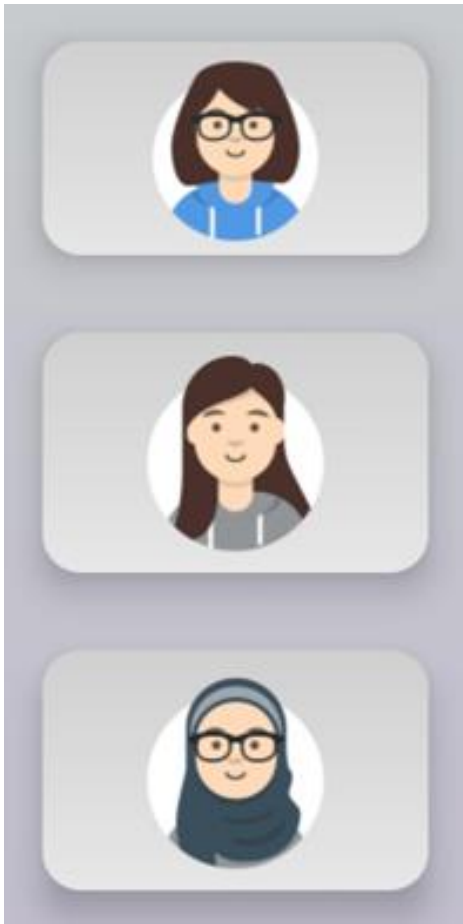


Figure 5.4: Choose female avatar

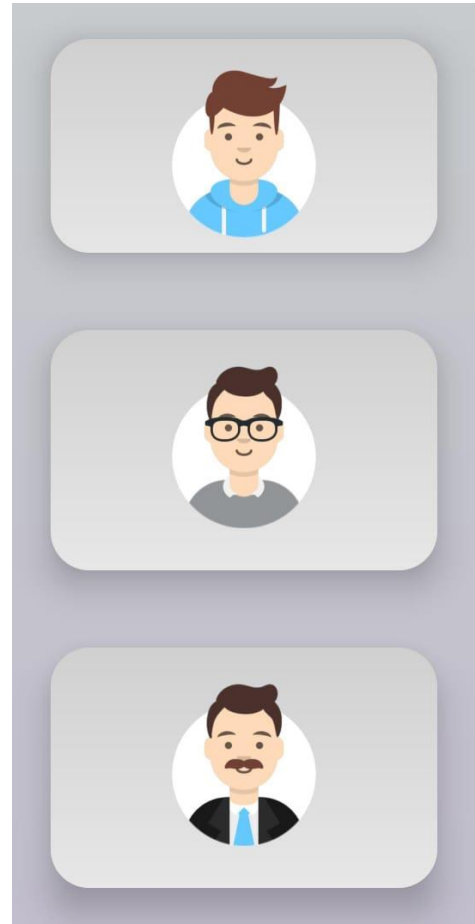


Figure 5.5: Choose male avatar

* Audio Upload and Recording Page:

- This page allows users to either upload an audio file from their mobile device or record their voice directly using the app. Once the audio input is provided, it sends a post request to upload the audio to API. When the call is successfully responded, users can generate a video where the selected avatar expresses the emotion detected in the speech. These steps are shown in **Figure 5.6, Figure 5.7, Figure 5.8.**
- In **figure 5.6**: the user clicks on the “Upload” button to access his mobile internal storage so he/she will go to **figure 5.7**.
- In **figure 5.7**: the user can choose to access his/her mobile internal storage then choose the audio in his/her mobile then the application will send post request to send the chosen audio to the API and send a Toast to user whether it was successfully responded or failed.
- **Figure 5.8 shows another option** where the user can click on record button and start talking then when he/she finishes, he/she clicks on it again then the application will send post request to send the recorded audio to the API and send a Toast to user whether it was successfully responded or failed. The user can play the audio by clicking on PLAY AUDIO button.

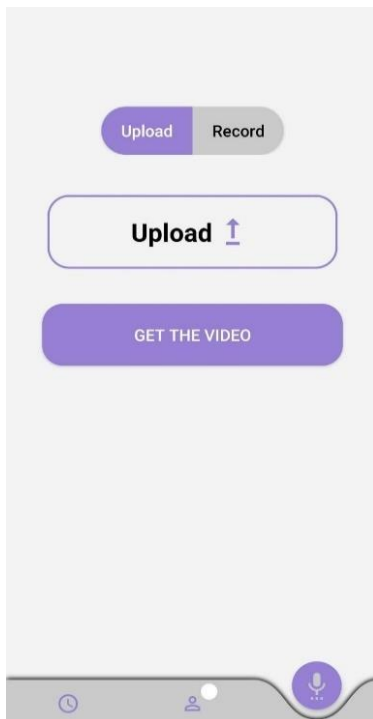


Figure 5.6: Upload audio page

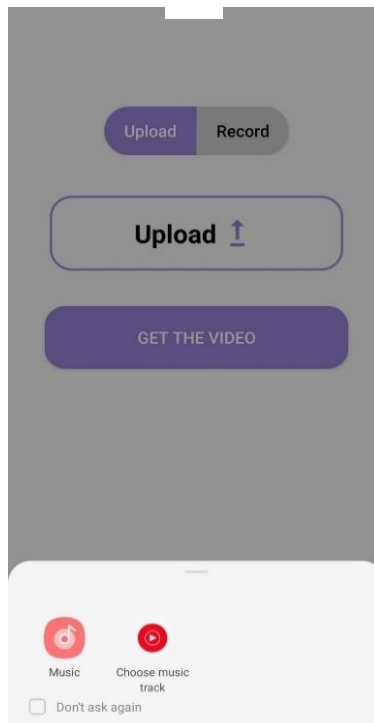


Figure 5.7: Allows users to choose audio file from their device to upload

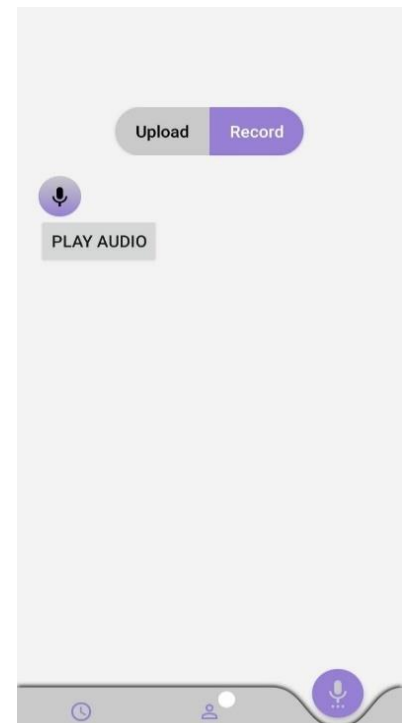


Figure 5.8: Record their voice using the app

* **Show video:**

- When the user finishes the above steps, they can click on the “GET THE VIDEO” button. When the button is clicked, a post request is sent that uploads the avatar number and the gender and runs the model to create the video on the specified audio. After successfully responding to it. A get request is sent to get the video created and is saved to the user’s mobile then shown for the user to play. The necessary steps are shown in [figure 5.9](#) and [figure 5.10](#).

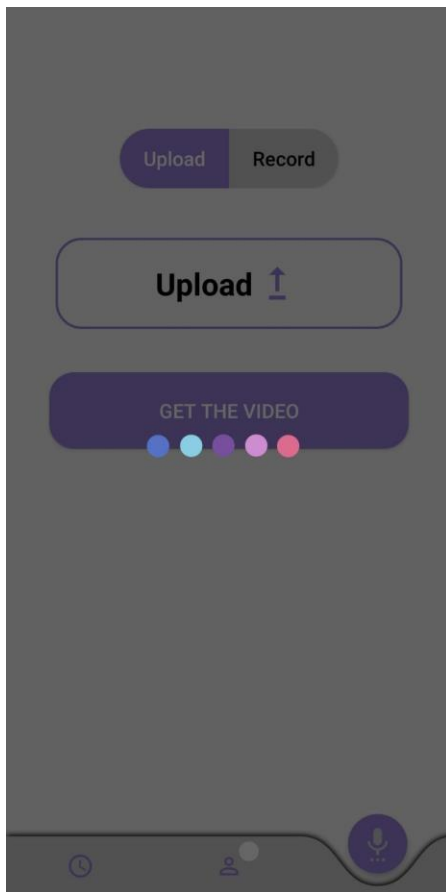


Figure 5.9: Waiting for creation of the video

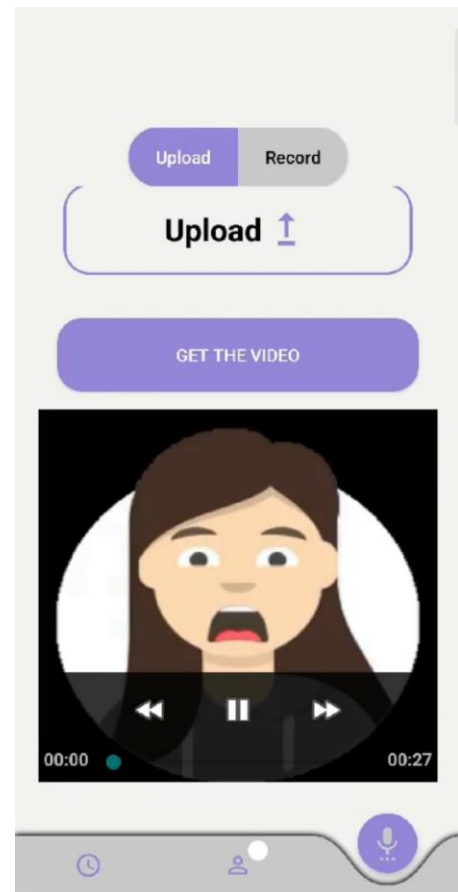


Figure 5.10: Show video

6- Conclusion and Future Work

6.1 Conclusion

In this project, our aim was to visualize emotions detected from speech using avatars. This is achieved by detecting emotions from speech using **SER** techniques and **Avatar generation** techniques to visualize the detected emotions.

For speech emotion recognition, we applied 2 models: CNN and Multi-Head Convolutional Transformer model and the LSTM model. We experimented with different configuration for both models and using many datasets such as RAVDESS [12], SAVEE [13], TESS [14] and even a combination between RAVDESS and TESS. Eventually, the **CNN and Multi-Head Convolutional Transformer**[8] model trained on the **RAVDESS** [12] dataset was used as our main SER trained model as it achieved the best accuracy compared to the other combinations of models with different datasets. We observed that the model best detects the distinct emotions of “angry” and “sad” and could not differentiate between “calm” and “neutral” emotions as they are close to each other in the voice tone. For visualizing emotions through avatars, we made suitable combinations of the avatar attributes (mouth, eyebrows,...etc.) so the generated avatars reflected the detected emotions, then we created 2 copies of each avatar (one with opened mouth and one with closed one) so we can imitate the visualization of speaking. Finally, a video of the set of the generated avatars was created.

6.2 Future Work

Possible venues for future work include:

1. Avatar personalization while preserving the identity of the original speaker and the addition of lip-syncing using deep fake technologies without the reliance on camera footage.
2. Visualizing the emotions of multiple users at the same time
3. Improving the accuracy of detection of the calm and the natural emotions for better differentiation between them.
4. Work on minimizing response time of the API.

References

- [1] J. Williamson, "Speech analyzer for analyzing pitch or frequency perturbations in individual speech pattern to determine the emotional state of the person ", Jun. 1978
- [2] S.T. Monisha and S. Sultana, "A Review of the Advancement in Speech Emotion Recognition for Indo-Aryan and Dravidian Languages", Dec. 2022, doi: 10.1155/2022/9602429
- [3] HT Bunnell and W. Idsardi, "Recognizing emotion in speech, Proceedings of the Fourth International Conference on Spoken Language Processing. ICSLP", vol. 1-4, Oct. 1996.
- [4] BS Kang, CH Han, ST Lee, DH Youn and C Lee, "Speaker dependent emotion recognition using speech signals, Proceedings of the Sixth International Conference on Spoken Language Processing", 6th International Conference On Spoken Language Processing (ICSLP 2000), Oct. 2000.
- [5] J. Fox and S. Ahn "Avatars: Portraying, Exploring, and Changing Online and Offline Identities", vol. 1, pp. 255-271, Jan. 2012, doi:10.4018/978-1-4666-2211-1.ch014
- [6] H. Attar, N. Kadole, O. Karanjekar, D. Nagarkar and S. More , "Speech Emotion Recognition System Using Machine Learning", International Journal of Research Publication and Reviews, vol. 3, no. 5, pp. 2869-2880, May 2022.

[7] N. Ristea, R. Ionescu and F. Shahbaz Khan, "SepTr: Separable Transformer for Audio Spectrogram Processing", Jun 2022.

[8] R. Ullah ,M, Asif ,W. Ali Shah ,F. Anjam, I. Ullah, T. Khurshaid 5, L. Wuttisittikulkij, S. Shah, S. Mansoor Ali and M. Alibakhshikenari, "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer", vol. 23, no. 13, Jul. 2023, doi:10.3390/s23136212

[9] H. Aouani and Y. Ben Ayed, "Speech Emotion Recognition with deep learning", vol. 176, pp.251-260, 2020

[10] H. Tang, D. Xu, N. Sebe and Y. Yan, "Attention-Guided Generative Adversarial Networks", Aug. 2019.

[11] IliaZenkov, "transformer-cnn-emotion-recognition", GitHub repository, 2020. [Online]. Available: <https://github.com/IliaZenkov/transformer-cnn-emotion-recognition/blob/main/README.md>

[12] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English", May 2018, doi:0.1371/journal.pone.0196391

[13] P. Jackson and S. Haq, "Surrey Audio-Visual Expressed Emotion (SAVEE) Database", Apr. 2015.

[14] K. Dupuis and K. Pichora-Fuller, "TESS", Jun, 1, 2010, [Online]. Available: <https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess>. [Accessed: Feb. 15, 2024].

[15] H. S. Kumbhar and S. U. Bhandari, "Speech Emotion Recognition using MFCC features and LSTM network", Jun. 2020, doi:10.1109/ICCUBEA47591.2019.9129067.

الملخص

"عند التعامل مع الناس، تذكر أنك لا تتعامل مع مخلوقات منطقية، بل مخلوقات عاطفية"

دليل كارنيجي

تشكل العواطف جانبًا أساسيًا من الإدراك البشري، وتؤثر بشكل عميق على التفاعلات الشخصية. ومع ذلك، فإن التحديات مثل الرهاب الاجتماعي يمكن أن تعيق الأفراد من التعبير عن عواطفهم بشكل فعال أثناء التواصل

هدفنا في هذا المشروع هو إنشاء تطبيق يسهل على الأشخاص التواصل مع بعضهم البعض وإظهار عواطفهم. لقد سعينا إلى تحقيق ذلك من خلال تصور عواطف شخص يتحدث - باستخدام الإشارات الصوتية فقط - من خلال إنشاء صورة رمزية مناسبة تعكس العاطفة المكتشفة من الكلام. يمكن تحقيق ذلك من خلال تطبيق التقنيات المتقدمة الحديثة في التعرف على عواطف الكلام

يمكن تقسيم إطار عملنا إلى وحدتين فرعيتين رئيسيتين: التعرف على عواطف الكلام وتوليد الصورة الرمزية. من أجل الوصول إلى أفضل نموذج للتعرف على عواطف الكلام، قمنا

بتطبيق تجارب مكثفة باستخدام نماذج ومجموعات بيانات مختلفة حتى تم تحقيق أفضل مزيج

مع SER كان هذا المزيج يستخدم نموذج CNN and Multi-head Convolutional

Transformer المدرب على مجموعة بيانات RAVADESS

والتي حققت دقة بنسبة 95%. يهدف هذا الإطار إلى تسهيل التعبير العاطفي للأفراد الذين قد

يواجهون صعوبة في التواصل اللفظي، مما قد يؤدي إلى ثورة في التفاعلات والتطبيقات

الشخصية في مجالات مختلفة. أخيرًا، يركز مشروعنا على تحسين التواصل البشري من خلال

تصور المشاعر من خلال الصور الرمزية الناتجة عن الكلام