Multimodal Speech and Facial Emotion Recognition System

ID Number	Name	Email
11597873	Ketha Tirumuru	kethatirumuru@my.unt.edu
11554879	Mani Deep Reddy Gadhe	manideepreddygadhe@my.unt.edu

Introduction:

In this project we are going to implement the Speech and Facial Emotion recognition system which is a multimodal system. Lets first discuss the Facial Emotion recognition system which is implemented using the model called Deep Convolutional Neural Network (DCNN). Here we are using the dataset called FER2013[3]. In this dataset they are having different emotions and all these images are labeled with images tag. In this dataset there are 6 labels. The labels of the images are sad, happy, anger, neutral, Disgust and Fear. Here we will use multiple Data Pre-processing techniques and prepare data for the modeling. Here we will use Neural Network with DCNN architecture. Here we use different augmentation techniques for the model to get different facial emotions[4]. This helps the model not go into the phase of overfitting. There are also different hyperparameters where we can make the model more accurate for the prediction. Some of the hyperparams include epochs, batch size, learning rates. After model building we need to have the performance metrics such as classification reports so that we can evaluate the model based on that report. If there are any disturbances then again we need to re-train the model by tuning the parameter tuning.

On the other hand we are also working on Speech emotion recognition. There, unlike FER we are considering multiple datasets. Datasets include RAVDESS, CREMA-D, TESS, and SAVEE. Here the most important library that we are using for audio is Librosa and we are utilizing keras framework to build the model[1]. Keras Framework is very efficient in extracting the insights from the speech signals. Here we have different steps for model preparation like Data preparation, data exploration and visualization using wave plots, spectrograms and emotion specific visualizations. The next step is we need to do data Analysis, Data Augmentation where we need to consider 3 types like noise, stretching, pitching and shifting. This helps the model not go into the phase of overfitting. Here we do Feature Extraction using Chroma_stft, MFCC, RMS value, Mel Spectrogram[2]. We need to one hot encoding for data normalization. There are also different hyperparameters where we can make the model more accurate for the prediction. Some of the hyperparams include epochs, batch size, learning rates,monitor, factor, verbose, patience, min_lr. After model building we need to have the performance metrics such as

classification reports so that we can evaluate the model based on that report. If there are any disturbances then again we need to re-train the model by tuning the parameter tuning.

Goals and Objectives:

- Goal: Here the main goal to do this project is to predict the emotion when the audio is provided and also to predict the emotion when the facial expression is given.

- Objectives:

1. **Develop Speech Emotion Recognition (SER) System:**
   - Speech Emotion recognition system which is implemented using the Kera Framework. Keras Framework is very efficient in extracting the insights from the speech signals. Which predicts the type of the emotion when the speech/audio file is given to the modal.

2. **Implement Facial Emotion Recognition (FER) System:**
   - Facial Emotion recognition system which is implemented using the model called Deep Convolutional Neural Network (DCNN). Which predicts the type of the emotion when the facial expression is given to the modal.

3. **Data preprocessing:**
   - By doing data preprocessing the imbalances in the data will be identified. We can apply dome of the data preprocessing techniques to reduce the imbalances in the data.

4. **Modeling:**
   - Here we need to create the models for both

5. **Modeling Performance:**
   - Here for both SER and FER we use the model performance metrics so that we will know how our model is performing on the new data. There are some metrics where we need to consider for model evaluation: Accuracy, Recall, Precision.

6. **Real-World Applications which uses these modals:**
   - The Real-World Applications include customer call centers where we use the speech/ audio from which we can detect the emotion and improve their service. The other real-world scenario where Drivers emotions will be captured and this helps to avoid accidents by capturing Drivers emotions and create an alert system.

Motivation:

The main motivation to develop this project or to select this project because this is very near to the practically real-time application because emotion is a very natural feeling where everyone has. Using that emotion and predicting something is a very useful real-time application. Here we used FER and SER which are natural.

As technology increases the importance of practical applications are increasing day by day. Suppose consider customer call centers where we use the speech/ audio from which we can detect the emotion and improve their service. The other real-world scenario where Drivers emotions will be captured and this helps to avoid accidents by capturing Drivers emotions and create an alert system.

Significance:

The SER and FER are one of the most important applications for today's real world scenario. These are used in different applications like customer call centers and FER is used in drivers emotion recognition while they are driving the vehicle. These are also used in different medical applications, SER and FER is one of the hottest Technologies in today's real world applications.

•Dataset:

Facial Emotion Recognition (FER) System:

1. fer2013.csv Dataset:

Now let's discuss the data sets used in SER and FER. Let's start with FER which is facial emotion recognition system. Here the dataset name which we used is FER 2013. CSV dataset[3]. This is the data set where we will be having different images With different emotions. These emotions are labeled means the images are given with the tag of the emotion. So this file consists of around 35,800 images. All these are official images with different emotions. In this data set there are six emotions considered which are 0 to 6 like angry discussed fear happy sad, surprise and neutral And each and every image will be having a pixel for each image will be having around 34,034 unique values for the pixels and the dataset is considered in such a way that 80% is considered for the training and 10% is considered for the testing the other 10% is considered for the other purposes. Example let's stay. Let's say they may be considered for the validation model validation gaining 10% for other purposes, including potential further testing or validation.

Speech Emotion Recognition (SER) System:

1. RAVDESS Emotional Speech Audio Dataset:

The next system we are considering is speech, emotion recognition. Here we consider data sets. The first data set is RAVDESS which is an emotional speech audio data set so as this is a speech recognition system, there will be audio files. These audio files are generated by the voice of actors, this dataset consists of 24 actors, 12 male and 12 female in total will be having 1440 files where they express different emotions through the speech. The emotions they consider are angry, fearful,

surprised, sad, happy and calm. The intensity in which they have given. These expressions are normally, strong and neutral[1].

2. CREMA-D (Crowd Sourced Emotional Multimodal Actors Dataset):
The next dataset we are considering for speech, emotion recognition is the CREMA-D data set. Here this data set consists of audio files generated by actors. The number of actors participating to generate these audio files are 91 in which 48 are males and 43 Are females, the age for the actors who have recorded these audios are between 20 and 74. In total we have 7442 audio clips which are generated by these actors and each actor has spoken 12 sentences in which they discussed fear, happy, neutral anger and sad emotions, here they have considered different levels of emotions like high, medium, low and specified.

3. Toronto Emotional Speech Set (TESS):
The next dataset which we have considered for speech, emotion recognition systems is the Toronto speech set. Here there are audio files which are recorded by the two female actresses. They have spoken 200 words and these actress are of age 26 and 64 the audio files which they have generated is in the format of WAV and the emotions which they have considered ouranger discussed, fear, happiness, present, surprise, sadness and neutral

4. Surrey Audio-Visual Expressed Emotion (SAVEE):
The SAVEE data set is the next dataset which we are considering for speech emotion recognition systems. In this dataset there are only male actors who have created audio files from their voice. The age of these male speakers are between 27 and 31. The emotions which we have created, our anger discussed, fear, happiness, sadness, surprise and neutral the number of speakers in this data set are four. This is completely different data set than other data sets because there are only male actors who are participated for the audio files generation.

Literature Survey:

[1]. In this paper they have considered the speed signals. They have created a model for speech, emotion, recognition classification. Here they have used auto encoder for selecting the parameters and the classification method which they have used is SVM but we have used neural networks which are more robust and more efficient than SVM.

[2]. In this paper they have considered speech signals and developed a multi model.

[3]. Here they have considered facial emotion recognition they have considered the same dataset which we have considered which is FERC2013. Here they have proposed the paper using convolution neural network as Deep Learning is very efficient for emotion recognition Here they have considered to data sets one is FRC, 2013 and Japanese female facial emotion

[4]. In this paper we have a foundation using neural networks on the FER 2013 data set given the adopted GGG Net architecture and the accuracy which they have obtained is 73%. We got 81% of accuracy which is more than this paper.

Features Analysis:

The features analysis is one of the parts of this project for building the model. This is one of the crucial steps where we extract the features and analyze them to get the relevant predictions for the emotions. In this project we are going to implement the Speech and Facial Emotion recognition system which is a multimodal system. Facial Emotion recognition system which is implemented using the model called Deep Convolutional Neural Network (DCNN)[1]. Here we will use multiple Data Pre-processing techniques and prepare data for the modeling. Here we will use Neural Network with DCNN architecture. Here we use different augmentation techniques for the model to get different facial emotions. This helps the model not go into the phase of overfitting. There are also different hyperparameters where we can make the model more accurate for the prediction. Some of the hyperparams include epochs, batch size, learning rates.
On the other hand we are also working on Speech emotion recognition. We are utilizing keras framework to build the model. Keras Framework is very efficient in extracting the insights from the speech signals. Here we have different steps for model preparation like Data preparation, data exploration and visualization using wave plots, spectrograms and emotion specific visualizations. The next step is we need to do data Analysis, Data Augmentation where we need to consider 3 types like noise, stretching, pitching and shifting. This helps the model not go into the phase of overfitting. Here we do Feature Extraction using Chroma_stft, MFCC, RMS value, Mel Spectrogram.

Methodology of SER:

1. Importing Libraries :

A. Pandas and NumPy:

- The library pandas are one of the important libraries for the data, manipulation and the analysis of the data. The pan has panda consist of data frame where it is a structured format for the data to be stored and this data can be used anywhere as as it is in the structure format
- Numpy one of the most important libraries for numerical operations. All the numerical operations are efficiently done through Namai and the storage of the Number is also very easy.

B. OS and SYS:

- The OS module is used when we are dealing with the parts folders et cetera in the python code.
- The system module is one of the important libraries in which we can interact with the python interpreter. In our case we use this library to eliminate the warnings when we execute the python.

C. Librosa:

- Librosa is one of the most important python libraries which deals with audio files. Using this library, we can extract the audio files. We can analyze some of the audio files. We can even extract the features from the audio files. Using this python library called Librosa.

D. Seaborn and Matplotlib:

- Seaborn is one of the visualization libraries in python. This will be working on the top of Matplotlib. It is used to create plots.
- Matplotlib is one of the complex visualization library, use it further plots and visualization of the data.

E. StandardScaler, OneHotEncoder, Confusion Matrix, Classification Report, Train-Test Split:

- Here the standard scale is generally used for the pre-processing. It removes the main and it will be scaling the variance.
- one card one hot encoding is used when we have a categorical target variable. This converts the targeted categorical variables into the matrix. This matrix is a binary matrix
- Confusion Matrix and the classification report this will be having a report where it gives the model performance like accuracy, recall precision and F1 score.
- By this performance metrics we can say how the model is performing. If it is not performing as expected we can find tune or hyper para so that we will be training the model and getting the expected predictions for the given data testing is very crucial. While we are building a model.
- The data is generally divided into train and test where the train data set is used to train the model and test data set is used to test the model.

F. IPython Display:

- Python display is used to display the audio files in the Jupyter notebook directly so that we can play the audio files in the Jupyter notebook

G. Keras:

- KERAS is one of the most efficient frameworks to build neural networks. This is built or running on the top of tensorflow.

2. Data Preparation:

A. Ravdess Dataframe:

- In the data preparation, the first data set we are considering is RAVDESS data frame
- First we will be loading this data set into a variable and this will be rated through all the files in the folder.
- These are the audio files which are generated by the actors. These audio files consist of information about the emotion, the intensity of the statement which they have repeated and the actor name.
- Here we will be having two variables in which one is stored for the emotion. The other variable stores the file path, the name of the variable, which to the emotion is emotion_df and variable name which stores file path_df in these data frames are combined to form ravdess_df and these data set emotions are converted into tax format so that we can read what is the text for the emotion.

B. Crema DataFrame:

- Crema data set as we have done for the above data set here also will be having a folder for the Crema data set where all the audio files for this specific data set will be stored here.
- Also we will be having two df. One with the list consists of the emotions called emotion_df and the other variable is called path_DF.
- Where the path of the file is stored. The two data frames are combined to form a single data frame called as crema_df frame and in this one the emotion is converted from the shortcut into the original text. Suppose ANG is converted into anger.

C. TESS DataFrame:

- TESS data frame the audio file is related to the data set will be stored in TSS directory.
- Now we will be having an iterative process where it goes through all the files in this data set and extracts the emotions and file parts in different variables, the emotions are stored inemotion_DF and the file path is stored in path_DF.
- These two data frames are combined into single data frame called as TESS__DF

D. CREMA-D DataFrame:

- SAVEE data frame, the audio files related to this data set will be stored in SVEE directly.
- Now we will be having an iteration process where it goes through all the files in the data set and next track, the emotions and file paths in different variables.

- The emotions are stored in emotion_DF and file path is stored in path_DF. These two data frames are combined into a single data frame called SAVEE_DF here, the emotion which is used as a prefix. For example, if they have considered ‘a’ then it is anger.

E. Combining Data Frames:

- Now, all the above data frames are combined into a single data frame and converted into CSV, the name of the CSV is data_Path.CSV

Data Exploration and Visualization:

A. Count of Emotions:

- Here we have considered that the count generally counts the number of unique values in the data. In our case in the X axis we are having different types of emotions and in the Y axis, We are having the count of each emotion so the count plot has plotted histograms for each emotion.

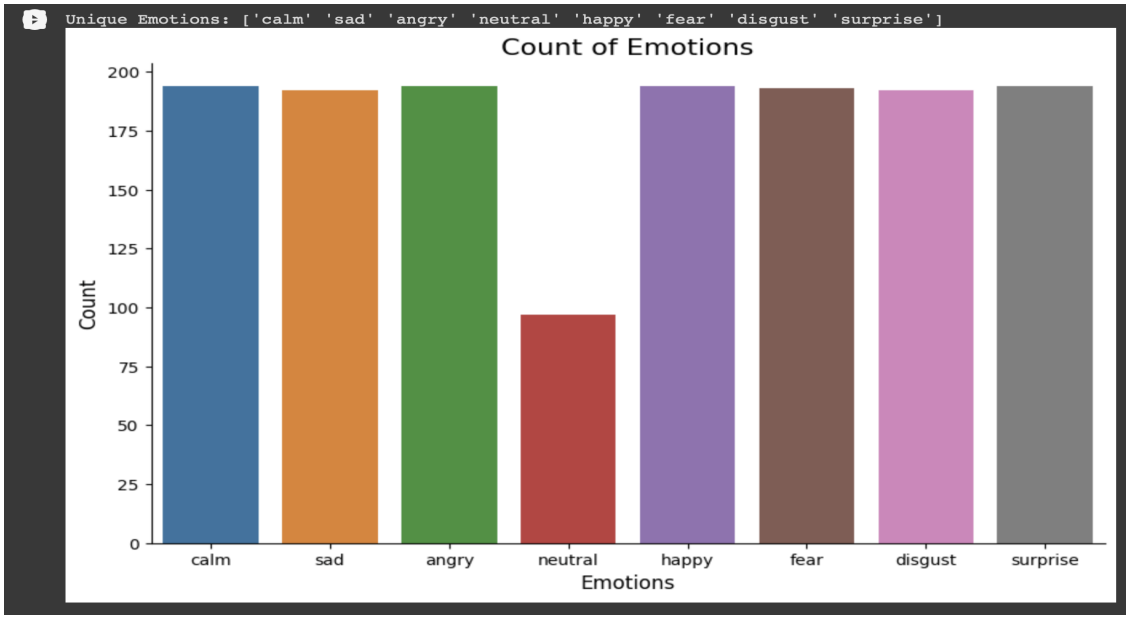


Fig1: Count plot for emotions

B. Wave Plots:

- Wave plots are under a kind of visualization where it visualizes waves of the different audio files with different emotions like fear, sad. Happy, anger

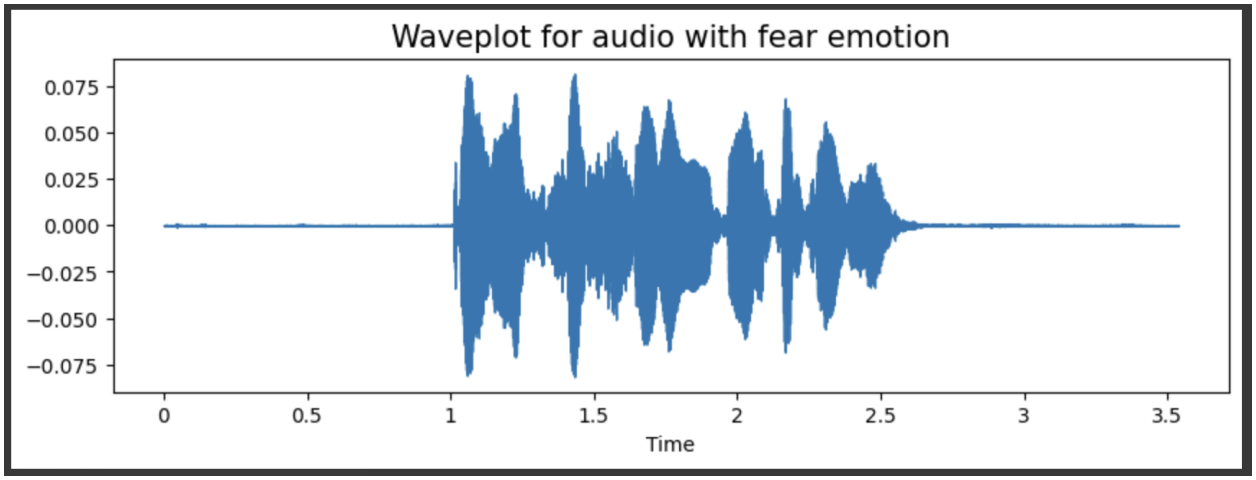


Fig 2: Waveplot for audio

C. Spectrograms:

- Spectrogram are the visualization created using the frequency of the audio signal where different audio signals are having different frequencies and color in the image shows that spectrogram represents the amplitude of different frequencies at different points of time.

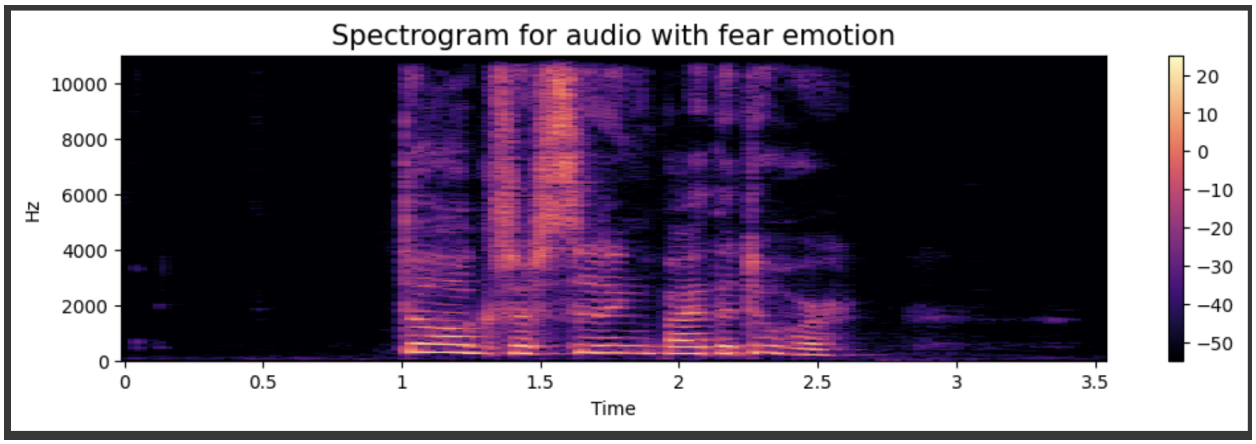


Fig3: Spectrogram for audio

D. Audio Playback:

- We will be having the audio file for each and every emotion. When we call the emotion type in the core it will be played in the Jupyter notebook directly. This is done using python library here using this audio playback we can connect to the audio and we can see the visual representation of the audio in the Jupyter notebook

Data Augmentation:

A. Noise Injection:

Data argumentation data argumentation is one of the important step for creating different audios from the original audio. Suppose in this case we are considering noise injection noise injection is a mechanism where we add external random voice to the original audio so that the data is argued with the noise[1].

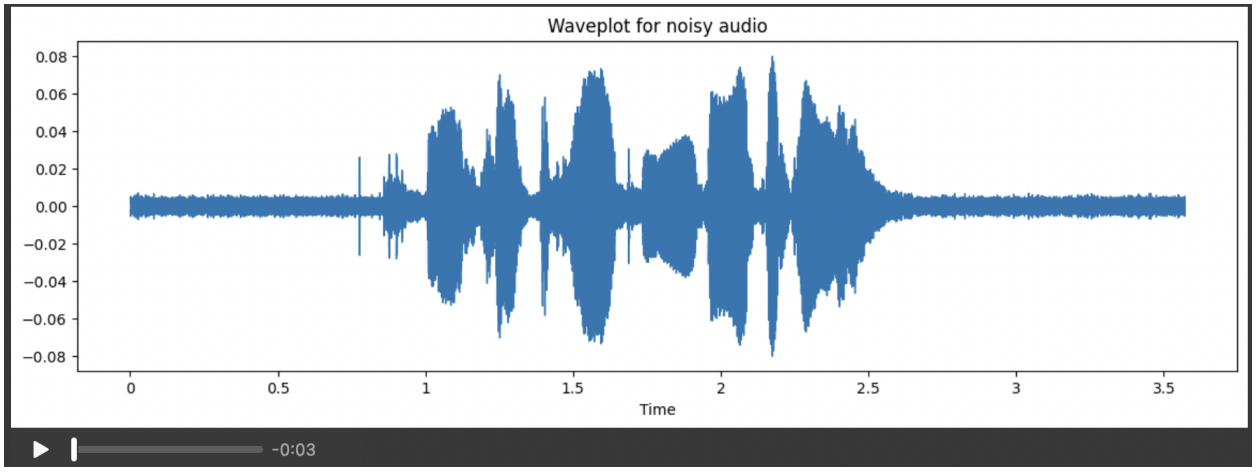


Fig4: Waveplot for noisy data

B. Time Stretching:

The other important data argumentation method is time stretching. This is also very important argumentation where it creates audios with different time. Duration is here will be stretching an audio for the different times without changing any pitch of the audio. For this one will be using a function called a stretch[2].

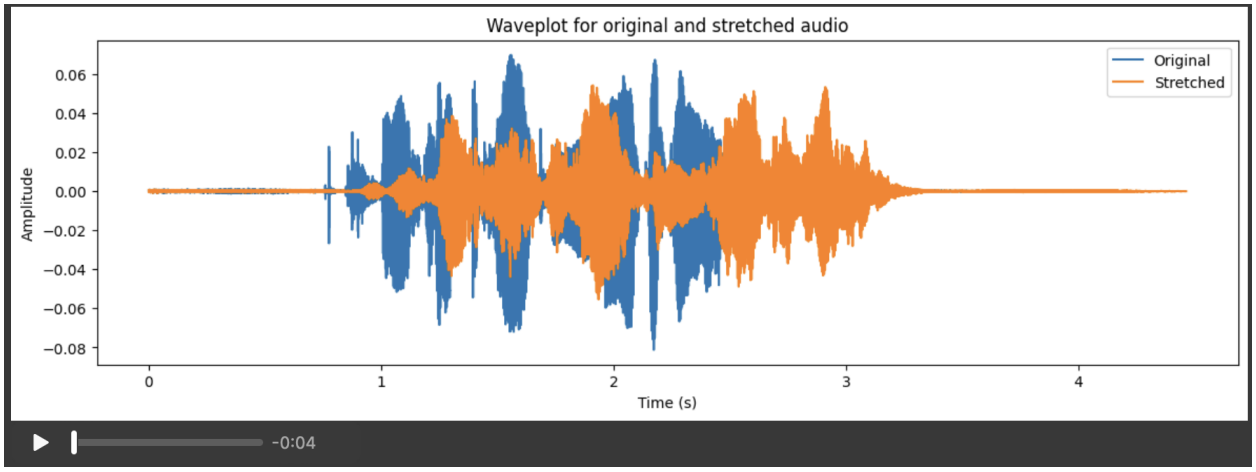


Fig5: Waveplot for stretch data

C. Shifting:

The other important data argumentation method for audio files is shifting shift is a function which is used to create this data argumentation here this moves the audio signal along its axis. The axis which it moves is called time axis. It will be having different variations in the timing and thespeech data, it will be having different types and variations in the rhythm of the speech.

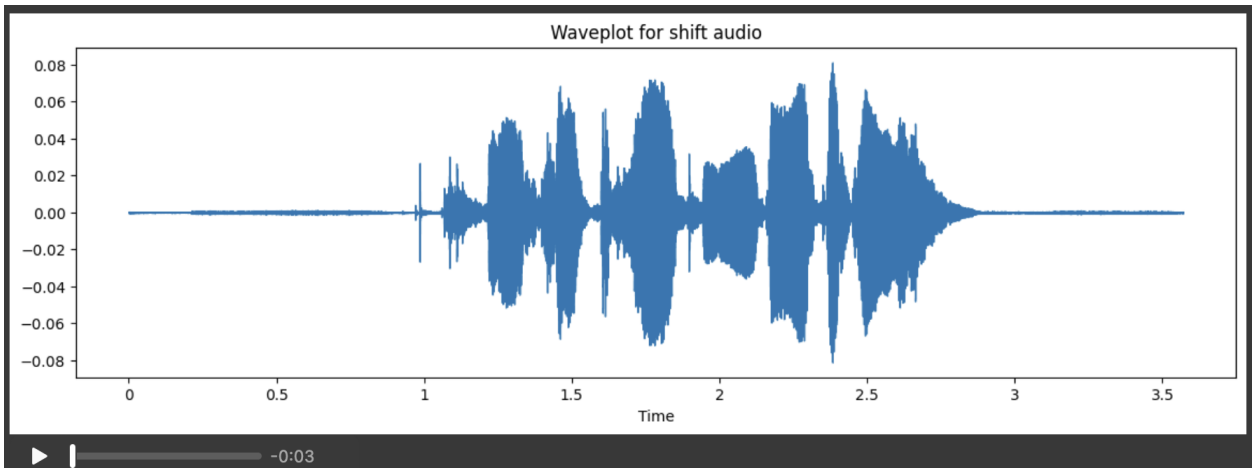


Fig6: Waveplot for shift data

D. Pitch Modification:

The other type of data argumentation for the audio signals is pitch. Unlike the above argumentation here will be changing the pitch of the audio signal or the audio file without changing the timing of the audio. The function which is used for this one is pitch.

All these above data argumentation is help us to remove the data imbalances because argumentation creates different data so that data balance will be detected and reduced. Then the model will be more robust for different data.

4. Feature Extraction:

Feature extraction is one of the most crucial step in any of the mission running projects mainly in this project, the audio files will not be understandable by the mission running model so these features should be transformed into mission understandable features

Data Preparation and Data Splitting:

A. Extracted Features:

The features need to be extracted and it stored in the data frame. Here we will be having X and Y where X contains all the labels columns in the data frame. Why contains all the columns which is nothing, but the targeted variable here in the files, the targeted variable is emotion.

B. One-Hot Encoding:

The audio files target variable is emotion. So here we need to apply the one hot encoding where it converts the categorical variable into the binary matrix. This metrics is useful for the training of the model as it is a multi class classification problem we need to use the one hot encoding[2].

C. Data Splitting:

Data speeding for any mission learning models, we need to split the data into training and testing generally will be taking 70% as training data and 30% as testing data. This can be done through a function which is imported from the sky called as train test and X, Y labels are divided into X and white train and testing represent represent and test

D. Standard Scaling:

In the audio signals, performing standard scaling is very important. In standard scaling we need to maintain a mean of zero and standard deviation of one. This is mainly done to prevent the domination of one feature on the other feature[1].

E. Data Preparation:

Data preparation is also one of the crucial and important steps while creating a convolution neural network. Here we need to convert it into a three dimensional array means we need to add extra dimension which is done through expand_DMS

Final Data Shapes:

- After preparation, the shapes of the data arrays are as follows:
  - `x_train`: (27364, 162, 1) - 3D array for training features.
  - `y_train`: (27364, 8) - 2D array for training labels (one-hot encoded).
  - `x_test`: (9122, 162, 1) - 3D array for testing features.
  - `y_test`: (9122, 8) - 2D array for testing labels (one-hot encoded).

Modeling:

A. Model Architecture:

Model architecture is very important. When we start building a model here we are creating a model using Keras, this is built on the top of Tensorflow. This architecture will consist of a 1D convolution layer. This will be having an active layer called as and it is a multi class classification. So we use the output layer activation function as soft Max activation. Here there several one day convolution layers are followed by the Max pulling dropout and then the dense layer[2].

B. Model Summary:

Model summary is the information in which the model will be trained on a number of parameters. In this case the total number of parameters on which the model is training is 557288.

C. Reducing Learning Rate on Plateau:

Here, the learning rate will be reduced using a reduce `ReduceLROnPlateau`, which is the callback given by the Keras

D. Training the Model:

Training the model is the crucial part in the modeling. Here the fit method is used on the and X test where the model is trained.
Training the model is a crucial part in modeling. Here If method is used on the train and train the model is strain based on this training data set here we considered 70% as a training data set and 30% as a testing data set. Now we have trained the model With 50 approx and the bad sizes 64 or model is validated based on the validation data set which is nothing but test and test.

E. Training Results:

After the model is trained we will get the training results. Nothing what the performance of the model we can get this one using accuracy and the loss.

F. Prediction and Evaluation:

Model evaluation is one of the important task in model building because in the model evaluation, we will get to know how the model is performing and the test rate here the predictions are made on test here previously we have used one hot and coding so we need to inverse those one hot and coding predicted labels into normal label so that we can know how many are predicted correctly and how many are predicted wrongly.

G. Confusion Matrix and Classification Report:

Confusion metrics are a classification report, which tells us the model performance like accuracy, precision, recall, and score for each one of the class. Here we can generate confusion metrics.

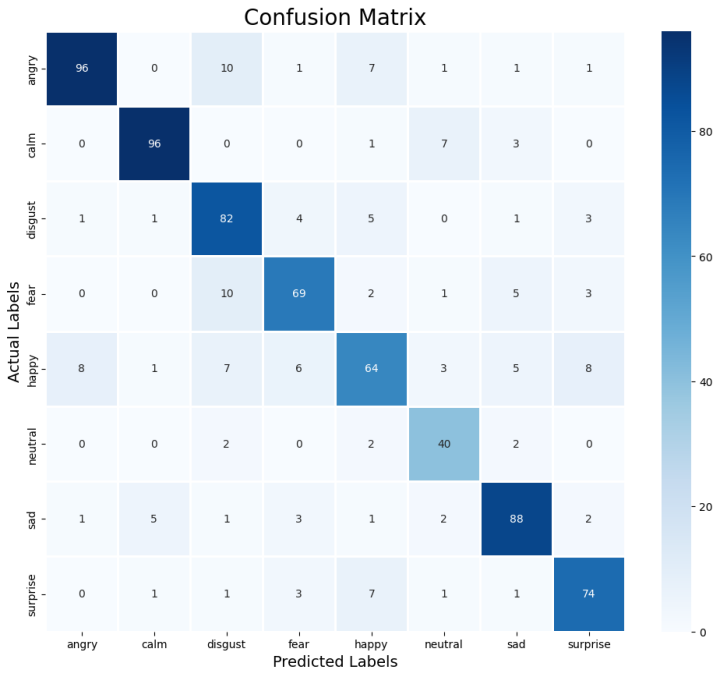


Fig7: SER confusion Matrix

	precision	recall	f1-score	support
angry	0.91	0.82	0.86	117
calm	0.92	0.90	0.91	107
disgust	0.73	0.85	0.78	97
fear	0.80	0.77	0.78	90
happy	0.72	0.63	0.67	102
neutral	0.73	0.87	0.79	46
sad	0.83	0.85	0.84	103
surprise	0.81	0.84	0.83	88
accuracy			0.81	750
macro avg	0.81	0.82	0.81	750
weighted avg	0.81	0.81	0.81	750

Fig8: Classification Report

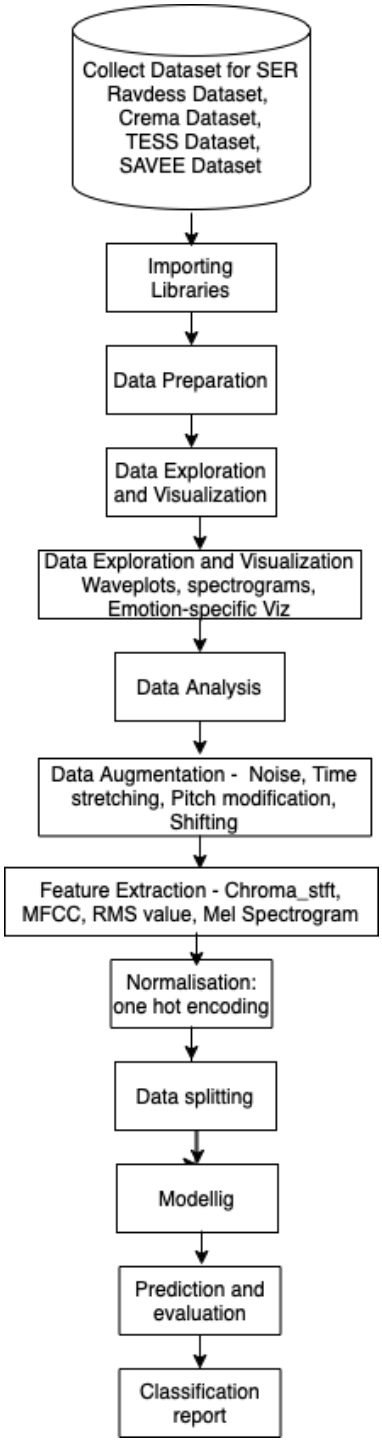
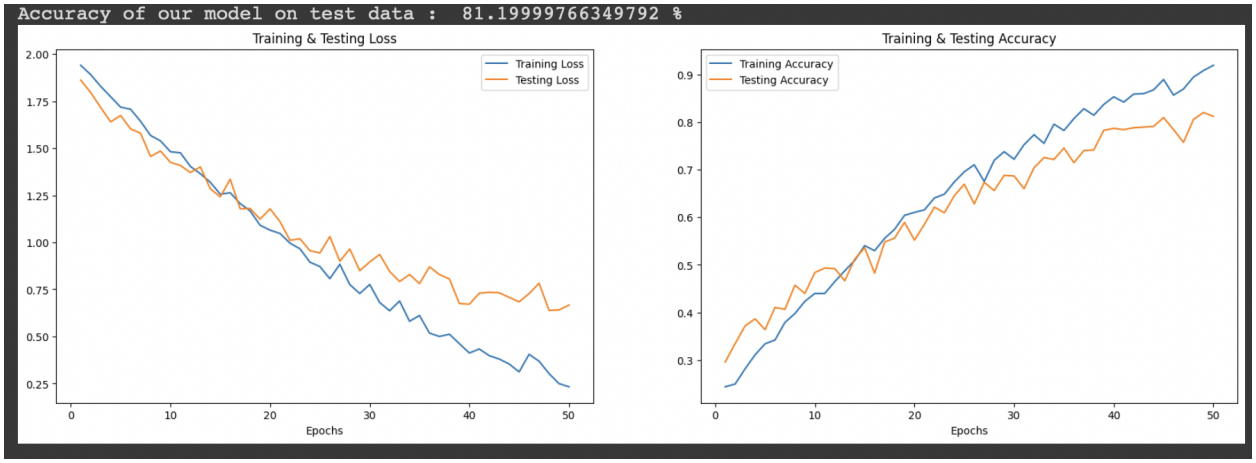


Fig9: SER Architecture

Results/Model Evaluation:

This graph represents the improvement in accuracy and decrease in the loss.



Model evaluation:

	Predicted Labels	Actual Labels
0	sad	sad
1	disgust	disgust
2	surprise	surprise
3	calm	calm
4	calm	calm
5	angry	angry
6	angry	angry
7	calm	calm
8	calm	calm
9	sad	sad

Facial Emotion Dataset (FER):

Dataset Overview:

In facial emotion recognition data set we have considered the data set called FER2013. And the data consists of 35,887 rows and three condoms. The three columns are emotion, pixels and usage.

Dataset Columns:

1. Emotion:
It is a numerical feature where emotions range from 0 to 6 pixels.

2. Pixels:
These are gray care in intensities usage how we use the record,

3. Usage:
whether it is used for training, testing or validation

Class Distribution Visualization:

Here, the code uses matplotlib to bring the counter plot. The majority of the emotions are happy, sad and neutral.

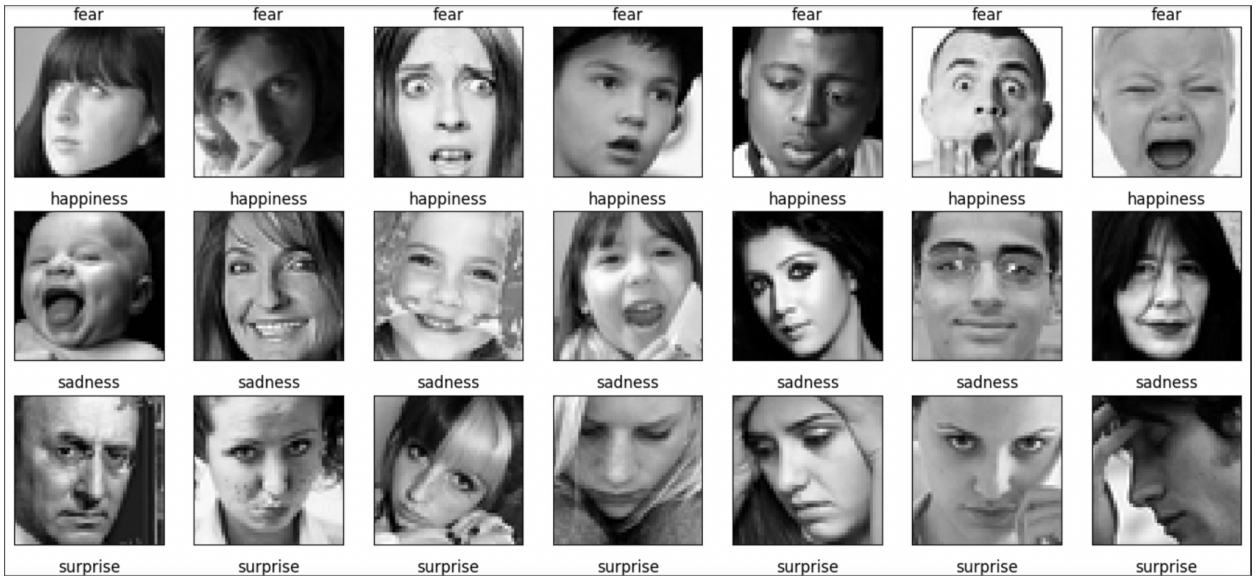


Fig10: Different Facial emotions

Preparing Data for Neural Networks:
The pixel values are extracted from the pixel column. Here, the column is pixel to 4 dimensional array and then we use a label encoder for converting the data into one hot encoding

Class Mapping:
Now we are having a dictionary which contains the names of the emotions corresponding to the value which is encoded

Data Splitting:
The next step is data, splitting, splitting the data into train and test.

Data Dimensions for Neural Network:
Here variables represent the dimensions of the data set meaning the number of classes increases the data set complexity increases and the building of neural network is also a bit complicated

Data Normalization:
In the data normalization, we take the values of the pixel from the training and validation to normalize the range to 0 to 1. This can be done by dividing each pixel value by 255. This is a very crucial step as neural networks are sensitive to unarmed data.

•Implementation status report

▪Work completed:

•Responsibility(Task,Person)

1. Data collection-4 datasets(SER) - Both of us
2. Data Preparation(SER) - Manideep
3. Data Exploration and Visualization(SER) - Manideep
4. Data Analysis(SER) -Manideep
5. Data Augmentation(SER) -Ketha
6. Feature Extraction(SER) -Manideep
7. Normalization(SER) -Ketha
8. Data splitting(SER) - Ketha
9. Model building/training(SER) - Ketha
10. Model Evaluation(SER) - Ketha
11. Classification Report(SER) - Manideep
12. Data collection(FER) - Both of us
13. Class distribution and visualization(FER) - Manideep
14. Preparing Data for Neural Networks(FER) - Ketha
15. Documentation - Both of us
16. PPT - Both of us

•Contributions (members/percentage):

1. Ketha Tirumuru - 50%
2. ManiDeep Reddy Gadhe - 50%

▪Work to be completed

•Responsibility(Task,Person)

1. Model building/training(FER) - Ketha
2. Model Evaluation(FER) - Ketha
3. Classification Report(FER) - Manideep
4. Need to make documentation more effective manner - Both of us
5. Try to improve Speech recognition Accuracy- Both of us

References/Bibliography:

[1]. Hadhami Aouani et.al.,“Speech Emotion Recognition with deep learning”, Volume 176, 2020, Pages 251-260.

[2]. Taiba Majid Wani et.al., “A Comprehensive Review of Speech Emotion Recognition Systems”, 22 March 2021

[3]. Akriti Jaiswal et.al., “Facial Emotion Detection Using Deep Learning”, 03 August 2020

[4]. Zhuofa Chen et.al., “Facial Emotion Recognition: State of the Art Performance on FER2013”, 8 May 2021

Github Link: https://github.com/tketha/Feature-Engineering-Emotion-Recognition