TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PURWANCHAL CAMPUS

[Subject Code: CT 707]

A PROJECT REPORT ON

"SPEECH EMOTION RECOGNITION USING RAVDESS, CREMA-D AND TESS AUDIO DATASETS & CNN AS DEEP LEARNING MODEL."

Submitted by:

Amrit Poudel **(PUR075BCT010)**

Anil Karki **(PUR075BCT011)**

Dilip Khadka **(PUR075BCT028)**

Jeevan Raj Panta **(PUR075BCT041)**

Submitted to:

Department of Computer and Electronics Engineering

Purwanchal Campus

Dharan, Nepal

A PROJECT WAS SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE BACHELOR'S DEGREE IN COMPUTER ENGINEERING

Date: March, 2023

ABSTRACT

Speech emotion is the technology that uses machine learning algorithms to analyze the emotion content of human speech. Speech emotion plays a prominent role in human-centered computing. This process involves extracting acoustic features from the speech signal and uses machine learning models to classify the emotion expressed in speech. However, acoustic recognition of emotion from the speech is a challenging task due to the complexity of emotional expression. The accuracy of emotion detection algorithms has steadily improved over time and is now capable of accurately detecting a range of emotions like: happiness, sadness, disgust, fearful, surprises, etc. As the field continues to develop, it holds great promise for improving human machine interaction and enhancing the understanding of human speech. In this work prosodic and spectral features are used for speech emotion recognition because both of these features contain emotion information. The potential features are extracted from each utterance for computational mapping between emotions and speech pattern. This project comprises researches, data collection, model building, testing, optimization and development of web based UI for prediction of input.

Keywords: Speech Emotion Recognition, Human Speech, Acoustic Recognition, Spectral Features, Computational Mapping, Speech Pattern.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
IDFT	Inverse Discrete Fourier Transform
LFPC	Log Frequency Power Coefficients
LPCC	Linear Prediction Cepstral Coefficients
MFCC	Mel-Frequency Cepstral Coefficients
MLR	Multivariate Linear Regression
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech
SER	Speech Emotion Recognition
SVM	Support Vector Machines
TESS	Toronto Emotional Speech Set

LIST OF FIGURES

1 INTRODUCTION

The fundamental purpose of speech is human communication; i.e., the transmission of messages between a speaker and a listener. Human communication is capable of conveying messages between each other with underlying emotions in it. Each speech signal has its own emotional attachment. However, only humans are capable of recognizing the attached emotions in speeches by listening to the tone, prosody, and other features of speech. But what if a computer has the ability to interact with human emotions? This is where our project emphasizes on.

Speech Emotion Recognition (SER) is the task of detecting and interpreting human emotions from speech signals. This technique includes processing of various audio speech files collected from RAVDESS, TESS and CREMA-D datasets are classified into 8 different emotions such as happy, sad, anger, calm, fear, surprise, disgust and neutral by computers.

This project mainly focuses on implementation of Deep Learning Models classified as CNN Algorithms that consists of mainly two subsystems i.e., Training/Learning System and Decision-Making System. The overall system consists of voice activity detection, speech segmentation, signal pre-processing, feature extraction, emotion classification and statistics analytics of emotion frequency. For feature extraction, we use the famous computation approach called Mel Frequency Cepstral Coefficients (MFCC).

1.1 Background

Speech Recognition is the technology that deals with techniques and methodologies to recognize the speech from the speech signals. Various technological advancements in the field of artificial intelligence and signal processing techniques, recognition of emotion made it easier and possible. It is also known as “Automatic Speech Recognition”. It is found that voice can be the next medium for communicating with machines especially when using computer-based systems.

Emotion recognition is a technology that aims to identify and interpret human emotions based on various inputs such as facial expressions, speech, physiological signals, and text. The recognition of emotions is an important part of human communication. The development of emotion recognition technology has been driven by advances in machine learning, computer vision, and natural language processing. These technologies enable computers to analyze and interpret human emotional states in real-time and provide appropriate responses. Speech Emotion Recognition (SER), Facial Expression Recognition (FER), etc. are the important forms of emotion recognition.

SER involves the analysis of speech signals to identify the emotional state of the speaker. The recognition of emotions from speech is a challenging task due to the complexity of speech signals and the variability in emotional expression among individuals. Speech signals contain various acoustic features such as pitch, intensity, and spectral characteristics, which can provide important information about the speaker's emotional state.

Researchers use various techniques to extract these features from speech signals, including Mel-frequency Cepstral Coefficients (MFCCs), pitch, energy, and formants. These features are then used to train machine learning models such as Support Vector Machines (SVMs), Random Forests, and Neural Networks to classify the emotional state of the speaker.

The performance of SER systems on the RAVDESS, CREMA-D and TESS dataset is typically evaluated using metrics such as accuracy, precision, recall, and F1 score. The highest performing models achieve accuracy rates of around 80-90%, indicating that SER is a promising technology but still has room for improvement. Nonetheless, SER is a promising technology with numerous potential applications that can benefit society in many ways.

1.2 Problem Statement

The problem statement of SER is to automatically identify the emotional state of speakers based on their spoken words, intonation, and other acoustic features. However, this is a challenging task due to the complexity of speech signals and the variability in emotional expression among individuals. Speech signals are highly variable and can be affected by factors such as accent, tone, pitch, and speed of speech. Moreover, emotional expression can be influenced by cultural and individual differences, making it difficult to develop models that can accurately recognize emotions across different languages and cultures.

Additionally, the classification of emotions can be ambiguous, and different emotions can be expressed in similar ways. For example, anger and fear may manifest in similar physiological responses, such as increased heart rate and respiration rate, making it difficult to distinguish between these emotions based on speech signals alone. Moreover, noise and other environmental factors can affect the acoustic properties of speech signals, making it difficult for models to accurately recognize emotions.

Overall, the problem statement of SER is to develop robust models that can accurately recognize emotions from speech signals, handle variability in emotional expression across cultures and languages, and handle noisy speech signals in real-world environments.

1.3 Objective

- To classify various audio speech files into different emotions such as happy, sad, anger, calm, disgust, surprised, fear and neutral.
- To improve the accuracy, efficiency, and effectiveness of human-machine interactions.

1.4 Project Features

- Emotion detection
- Robust
- Speech signal sampling
- Classifier training
- Feature extraction
- Preprocessing

1.5 Feasibility

1.5.1 Technical Feasibility

Our project is solely based on the input dataset of sound files and the extent of features that can be extracted from the RAVDESS, TESS and CREMA-D audio datasets, with a number of existing algorithms and techniques available for implementation. Due to the availability of speech datasets, existing algorithms and techniques, advances in speech processing technologies, availability of software tools, technical feasibility of succeeding our project is quite optimum unless there is some noise interference of noise in datasets.

1.5.2 Economic Feasibility

The availability of a normalized audio dataset called RAVDESS can be extracted from good sources free of cost and the project is based on the existing technology i.e. machine learning. Due to the availability of existing technology and resources our project is economically justifiable.

1.6 Applications and Scope

- Marketing and advertising
- Healthcare
- Customer satisfaction
- Gaming experience improvement
- Social media analysis
- Human robot interaction
- Stress monitoring

2 LITERATURE REVIEW

Speech emotion recognition (SER) is an important field of research that aims to enable machines to detect and recognize emotions from speech signals. Various researchers have aimed to develop such a system which could make human and computer interaction easy and feasible. Various researchers have aimed to develop such a system which could make human and computer interaction easy and feasible.

Nwe et al. [1] proposed a new system for emotion classification of utterance signals. The system employed a short time log frequency power coefficient (LFPC) and discrete HMM to characterize the speech signals and classifier respectively. This method classified the emotion into six different categories then used the private dataset to train and test the new system. In order to evaluate the performance of the proposed method, LFPC is compared with the Mel-frequency Cepstral Coefficients (MFCC) and linear prediction Cepstral coefficients (LPCC). Results demonstrate the average and best classification accuracy achieved 78% and 96% respectively. Furthermore, results show that LFPC is a better option as a feature for emotion classification than the standard features [1].

Automatic Speech Emotion Recognition Using Machine Learning-By Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, KosaiRaoof, Mohamed Ali Mahjoub and Catherine Cleder. In this paper, presents a comparative study of speech emotion recognition (SER) systems. To achieve this study, an SER system, based on different classifiers and different methods for features extraction, is developed. Mel-frequency Cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted from the speech signals and used to train different classifiers. Feature selection (FS) was applied in order to seek for the most relevant feature subset. Several machine learning paradigms were used for the emotion classification task. A recurrent neural network (RNN) classifier is used first to classify seven emotions. Their performances are compared later to multivariate linear regression (MLR) and support vector machines (SVM) techniques, which are widely used in the field of emotion recognition for spoken audio signals.

3 REQUIREMENT ANALYSIS AND SPECIFICATION

3.1 Functional requirements

- The system should be able to take audio input from the microphone.
- The system should be able to determine the emotion correctly.
- The system should be able to classify the emotion of the speaker.
- The system should be able to process the audio input in real time.

3.2 Non-functional requirements

- The system should be able to process audio input quickly and efficiently, with minimum delay or lag.
- The data content should be secured properly.
- The system should be compatible with various operating systems, programming languages and hardware.
- The system should be able to handle large volumes of data.

4 METHODOLOGY

4.1 Project Timeline

The timeline for our project is shown in figure 4.1;

Activity/Month	First	Second	Third	Fourth	Fifth	Sixth
Research						
Model Development						
Model Implementation						
Testing and Debugging						
Output Analysis						
Documentation						

Fig 5 .1 Gantt chart of Project Development

4.2 Tools

4.2.1 Python

Python is a popular high-level Programming language that is widely used for a variety of purposes, including web development, data analysis, artificial intelligence, scientific computing, and more. It was created by Guido Van Rossum in the late 1980s and was first released in 1991. Python has a design philosophy that emphasizes code readability and a syntax that allows programmers to express concepts in fewer lines of code that might be used in languages such as C++ or Java.

Python is derived from many other languages including ABC, Modula-3, C, C++, Algol-68, Smalltalk, and UNIX shell and other scripting languages. Python is copyrighted.

Python has a large and active community of developers who contribute to a vast array of libraries and frameworks that extend its capabilities. Some of the most popular libraries include NumPy, Pandas, Matplotlib, Librosa, Keras and TensorFlow.

4.2.2 NumPy

NumPy is a powerful and efficient library that provides essential functionality for scientific computing and data analysis in Python. Numpy can be used as an efficient multi-dimensional container of generic data along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of Numpy, Numeric, was originally created by Jim Hugunin with contributions from several other developers.

NumPy provides a powerful array processing functionality that is both efficient and convenient. It is built on top of the C programming language, which provides it with high performance. NumPy arrays are also more efficient than Python's built-in list data structure, especially for large datasets.

4.2.3 Pandas

Pandas is a popular open-source data manipulation and analysis library for Python. Pandas is widely used in data science and machine learning applications for tasks such as data cleaning, data exploration, data transformation, and data visualization. It also provides powerful capabilities for data aggregation, grouping, filtering, and merging.

Features

- Pandas provides powerful tools for cleaning and transforming data, including methods for handling missing values, data type conversions, and data reshaping.
- Reshaping and pivoting of data sets.
- Data alignment and integrated handling of missing data.
- Pandas provides tools for combining multiple data frames based on common columns or indices.

4.2.4 Librosa

Librosa is a valuable python music and sound investigation library that helps programming designers to fabricate applications for working with sound and music document designs utilizing Python. Librosa also includes functionality for feature extraction and signal processing, such as time-domain and frequency-domain filtering, and resampling. At a high level, librosa provides

implementation of a variety of common functions used throughout the field of music information retrieval. It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals and also do the feature extractions in it using different signal processing techniques like MFCC, RMS, chroma_stft , zero crossing rate, mel-spectrogram etc.

4.2.5 Keras

Keras is an open-source high-level Neural Network library, which is written in Python and is capable enough to run on Theano, TensorFlow, or CNTK. It was designed to enable fast experimentation with deep neural networks, it focuses on user-friendly, modular, and extensive. Keras provides a wide range of pre-built layers, including convolutional, recurrent, and dense layers, as well as activation functions and loss functions. It also provides tools for compiling and training models, including optimizers, metrics, and callbacks. It is made user-friendly, extensible, and modular for facilitating faster experimentation with deep neural networks. It not only supports Convolutional Networks and Recurrent Networks individually but also their combination.

4.3 Techniques

4.3.1 Data Acquisition

The first step is to collect a large dataset of speech samples that are labeled with the corresponding emotions. The dataset should include samples of speech for each emotion that you want to detect like sad, happy, neutral, angry, fear, surprise etc. For this project, the required datasets were acquired from KAGGLE. For this project we used the 4 different datasets and all of them were acquired from KAGGLE datasets. The 3 different datasets includes:

4.3.1.1 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Datasets:

Files

The RAVDESS is a validated multimodal database of emotional speech and song. This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. 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.

File naming convention

Each of the 1440 files has a unique filename. 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:

Filename identifiers

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). NOTE: There is no strong intensity for the 'neutral' emotion.

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).

Filename example: 03-01-06-01-02-01-12.wav

Audio-only (03)

Speech (01)

Fearful (06)

Normal intensity (01)

Statement "dogs" (02)

1st Repetition (01)

12th Actor (12) Female, as the actor ID number is even.

4.3.1.2 Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) is a publicly available dataset of audiovisual recordings of actors portraying different emotions. The dataset was created to aid research in the field of affective computing, which is the study of emotion recognition by computers. CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African American, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

4.3.1.3 Toronto Emotional Speech Set (TESS)

The Toronto emotional speech set, also known as the TESS, is a publicly available dataset of audio recordings of actors portraying different emotions. There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total. The dataset is organized such that each of the two female actors and their emotions are contained within its own folder. And within that, all 200 target words audio files can be found. The format of the audio file is a WAV format.

4.3.2 Data Augmentation

Data augmentation is a technique used in machine learning and computer vision to increase the size of a dataset by creating new data from the existing data. It includes making minor changes to the dataset or using deep learning to generate new data points. In sound speech recognition, data augmentation works wonders. It improves the model performance even on low-resource languages. The random noise injection, shifting, and changing the pitch can help you produce state-of-the-art speech-to-text models. You can also use GANs to generate realistic sounds for a particular application. The objective of data augmentation is to make our model invariant to those perturbations and enhance its ability to generalize. Numpy provides an easy way to handle noise injection and shifting time while Librosa help to manipulate pitch and speech.

4.3.2.1 Noise Injection

In noise addition data augmentation, we mix random value into data by using Numpy.

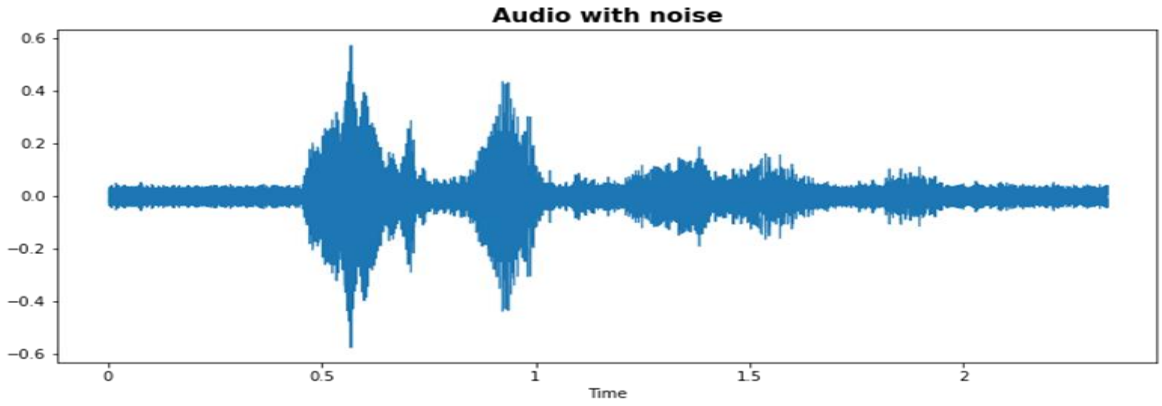


Fig.4.2 Audio with noise

4.3.2.2 Stretching

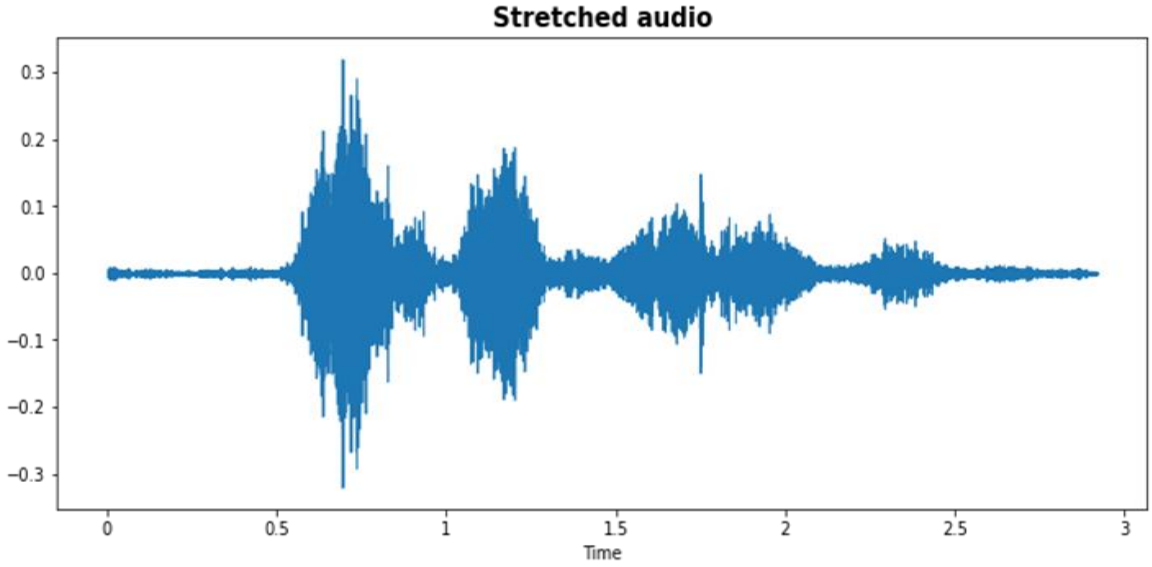


Fig. 4.3 Stretched Audio

4.3.2.3 Shifting

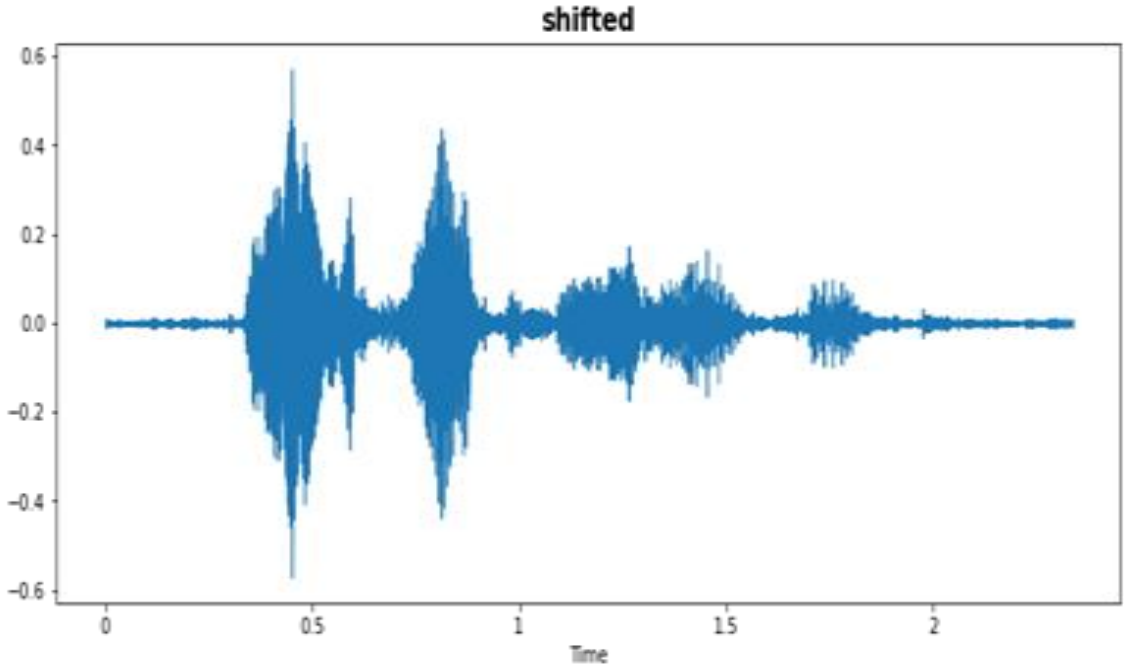


Fig. 4.4 Shifted Audio

4.3.2.4 Pitching

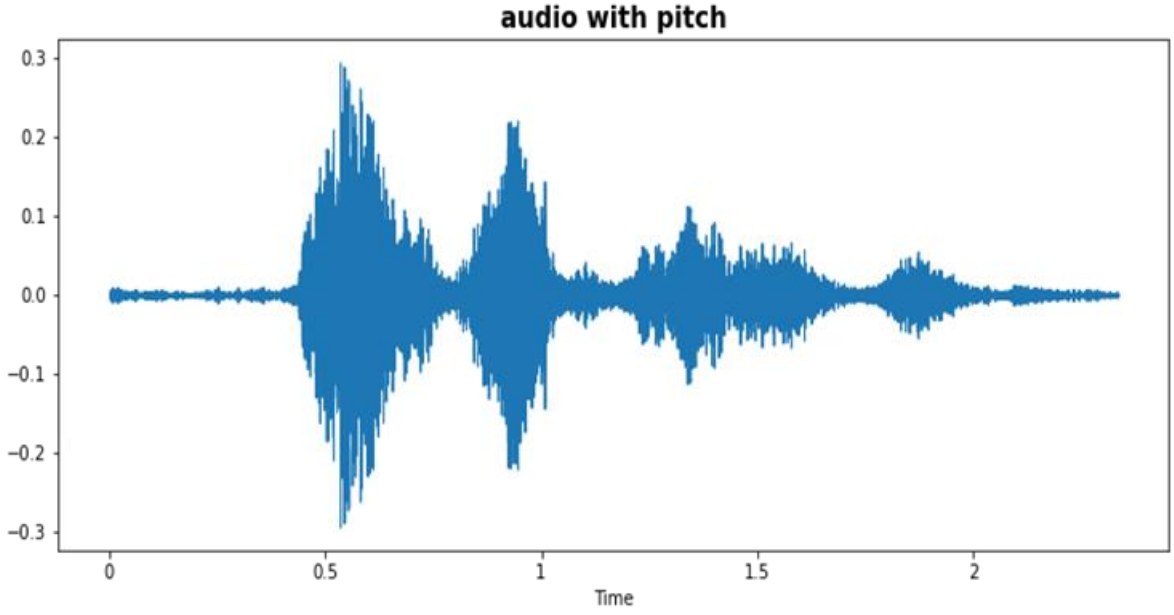


Fig. 4.5 Audio with Pitch

4.3.3 Feature Extraction

Features extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original dataset. The model cannot understand the provided audio data directly hence to convert them into understandable format the feature extraction is used. It is the process that explains most of the data but in an understandable way. Feature extraction in speech emotion detection involves selecting the most relevant acoustic features that capture the emotional content of speech and discarding irrelevant information that may be present in the raw speech signal. The various features extraction which we are using in our projects:

4.3.3.1 Mel Frequency Cepstral Coefficients (MFCC)

In the speech recognition problem, we can't take the raw audio signal as input to our model because there will be a lot of noise in the audio signal. It is observed that extracting features from the audio signal and using it as input to the base model will produce much better performance than directly considering raw audio signal as input. MFCC is the widely used technique for extracting the features from the audio signal. For a very basic understanding, cepstrum is the information of rate of change in spectral bands. Fig. 1 is the block diagram of the conventional MFCC extraction algorithm.

Features Extraction Using MFCC

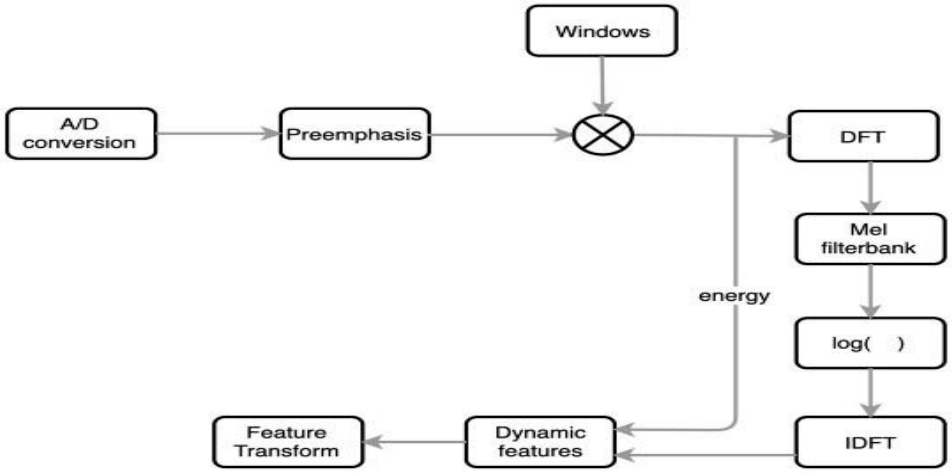


Fig. 4.6 Features Extraction technique Using MFCC

A/D conversion samples the audio clips and digitizes the content. A sampling frequency of 8 or 16 kHz is often used. The speech is first pre-emphasized with a pre-emphasis filter $1 - az^{-1}$ to spectrally flatten the signal, where "a" is between 0.9 and 1. Then the pre-emphasized speech is separated into short segments called sliding frames. The chopped frame with Hamming and Hanning maintains the original frequency information better with less noise compared to a rectangle window. The frame length is set to 20ms (160 samples) to guarantee stationary condition inside the frame.

$$Hamming\ (\alpha = 0.46164)\ \text{or}\ Hanning\ (\alpha = 0.5)\ \text{window}$$

$$w[n] = (1-\alpha) - \alpha\cos\left(\frac{2\pi n}{L-1}\right) \qquad L:\ \text{window width}$$

Next, we apply DFT to extract information in the frequency domain. The equipment measurements are not the same as our hearing perception. For humans, the perceived loudness changes according to frequency. Also, perceived frequency resolution decreases as frequency increases. In feature extraction, we apply triangular band-pass filters to convert the frequency information to mimic what a human perceives. We apply these triangular Mel-scale filter banks to transform it to Mel-scale power spectrum. The output for each Mel-scale power spectrum slot represents the energy from a number of frequency bands that it covers.

$$\boldsymbol{Mel(f) = 1127\ \ln\left(1 + \frac{f}{700}\right)}$$

Humans are less sensitive to small energy changes at high energy than small changes at a low energy level, which is logarithmic in nature. So our next step will take the log out of the output of the Mel filter bank. This also reduces the acoustic variants that are not significant for speech recognition. The log spectrum comprises information related to the phone and the pitch.

Our next step is to compute the cepstral which separates the glottal source and the filter. We can apply the inverse Fourier Transformation to separate the pitch information from the formants. After the IDFT (Inverse Discrete Fourier Transform), the pitch information with 1/T period is transformed to a peak near T at the right side.

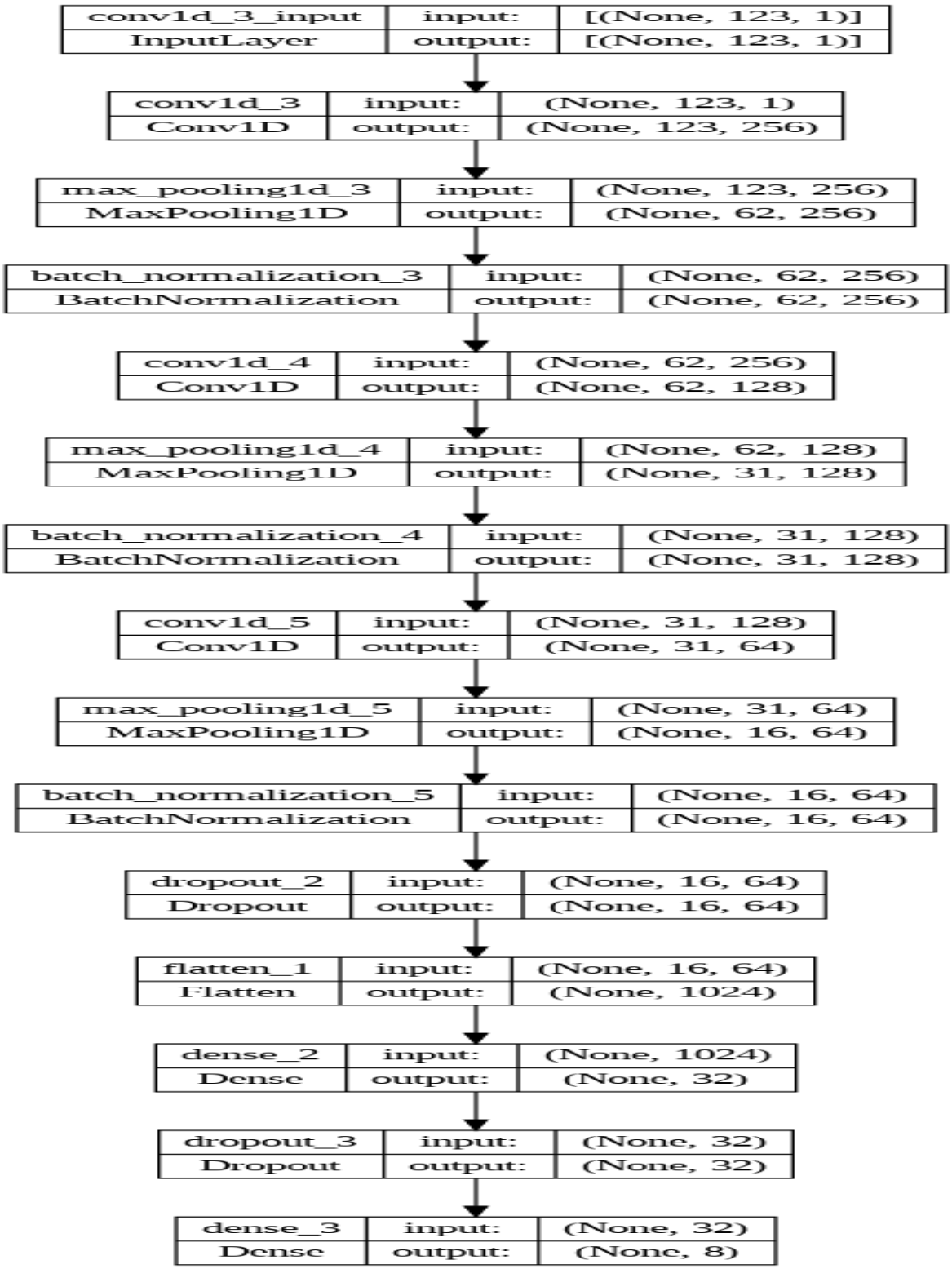
In fact, MFCC just takes the first 12 cepstral values. There is another important property related to these 12 coefficients. Its inverse DFT is equivalent to a discrete cosine transformation (DCT).

In ML, unrelated MFCC features make our model easier to model and to train. If we model these parameters with multivariate Gaussian distribution, all the non-diagonal values in the covariance matrix will be zero.

Lastly, MFCC has 39 features. We finalize 12 and what are the rest. The 13th parameter is the energy in each frame. It helps us to identify phones. Another 13 values compute the delta values. It measures the changes in features from the previous frame to the next frame. The last 13 parameters are the dynamic changes of delta from the last frame to the next frame. It acts as the second-order derivative. So the 39 MFCC features parameters are 12 Cepstrum coefficients plus the energy term.

The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum.

4.3.4 Modelling



4.4 System Design and Architecture

4.4.1 Software Development Approach

The meaning of Agile is swift or versatile. "Agile process model" refers to a software development approach based on iterative development. Agile methods break tasks into smaller iterations, or parts do not directly involve long term planning. The project scope and requirements are laid down at the beginning of the development process. Plans regarding the number of iterations, the duration and the scope of each iteration are clearly defined in advance. Each iteration is considered as a short time "frame" in the agile process model, which typically lasts from one to four weeks. The division of the entire project into smaller parts helps to minimize the project risk and to reduce the overall project delivery time requirements. Each iteration involves a team working through a full software development life cycle including planning, requirements analysis, design, coding, and testing before a working product is demonstrated to the client.

In our project we started with a few dataset using only two features extraction techniques. Gradually, after analyzing the result obtained from each iteration and comparing that result with our requirement we kept on iteratively increasing our datasets and feature extraction techniques until we got the satisfactory result. Hence our project was solely based on the agile development model.

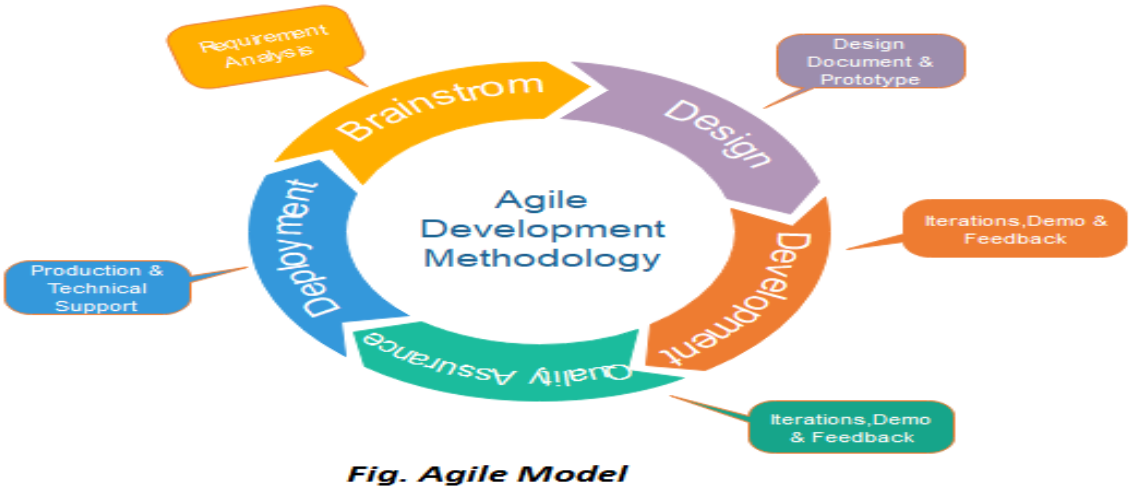


Fig. 4.7: Agile Software Development Model.

4.4.2 Overall System Design

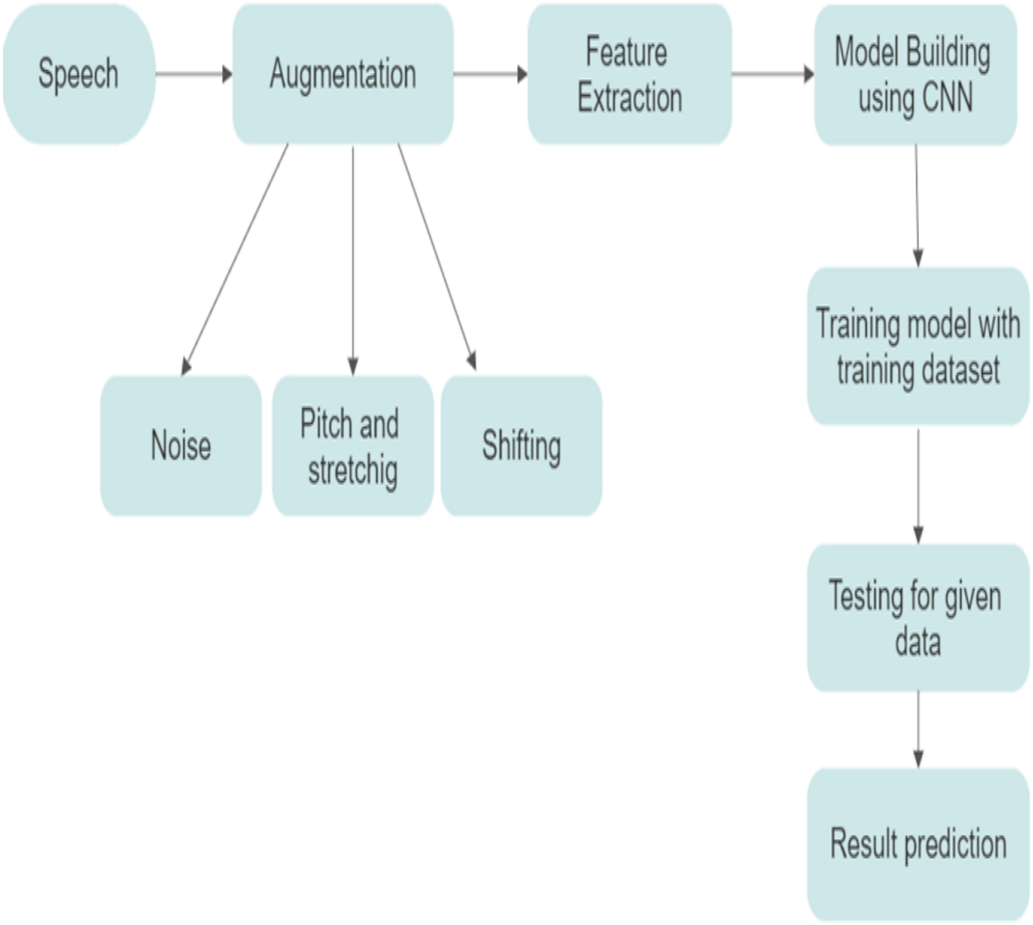


Fig.4.8: Overall System Design

5 SYSTEM OUTPUT

The accuracy of CNN model we trained for our project was found to be 79.27% which is capable of predicting the following emotions: happy, sad, fear, surprise, angry, disgust, calm and neutral. After feeding the input of our voice to the datasets externally, we found moderate change in accuracy and also the recognition was quite accurate for all three types: without noise, with noise and with pitching and shifting. The snapshots of the output of speech recognition system we built is given below:

5.1 Classification Scenarios

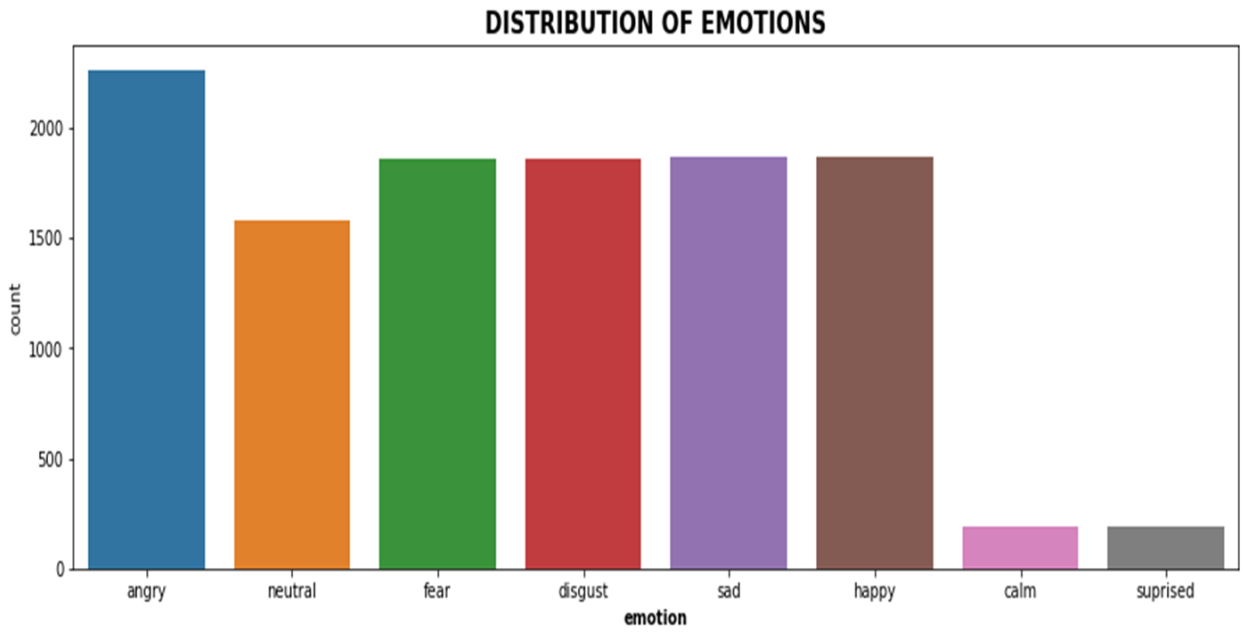


Fig.5.1: Distribution of Emotions

5.2 Expected Output (Actual Labels vs. Predicted Labels)

	Predicted Labels	Actual Labels
0	disgust	happy
1	happy	happy
2	happy	happy
3	disgust	fear
4	angry	angry
...	...	...
11692	neutral	neutral
11693	angry	angry
11694	disgust	disgust
11695	sad	sad
11696	disgust	disgust

11697 rows × 2 columns

Fig. 5.2: Output of Test Dataset

5.3 Training vs. Validation Loss & Accuracy Curve

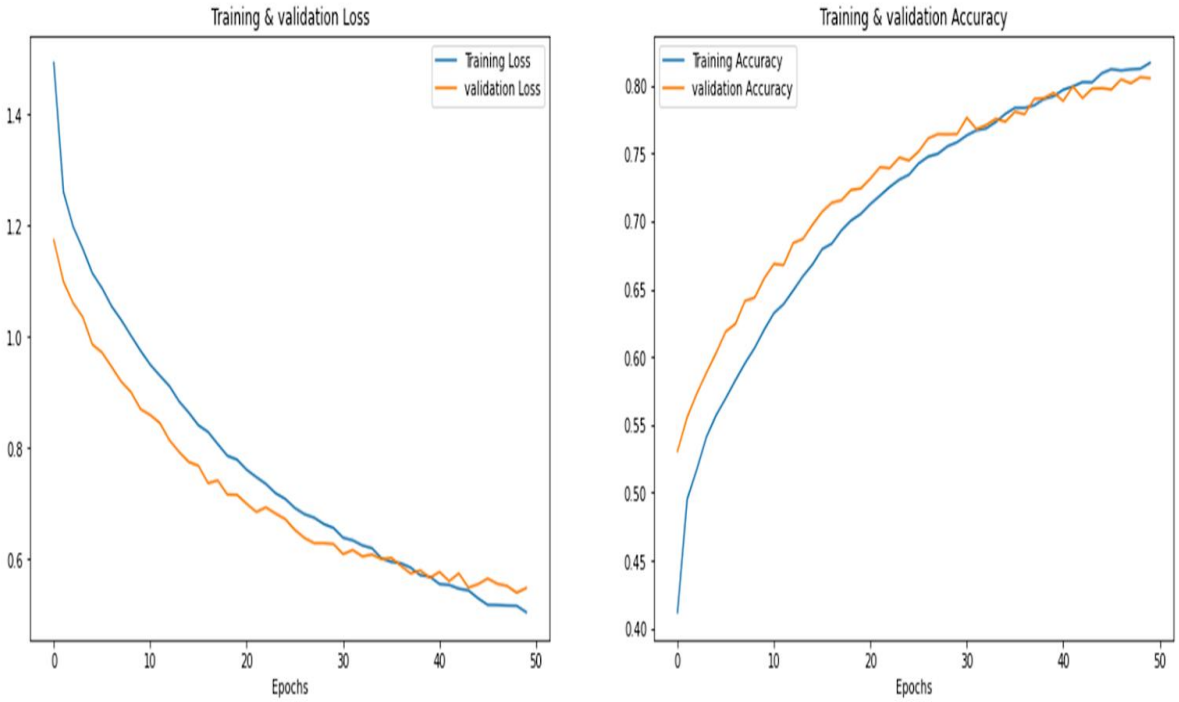


Fig. 5.3: Training and Validation Accuracy

5.4 Confusion Matrix

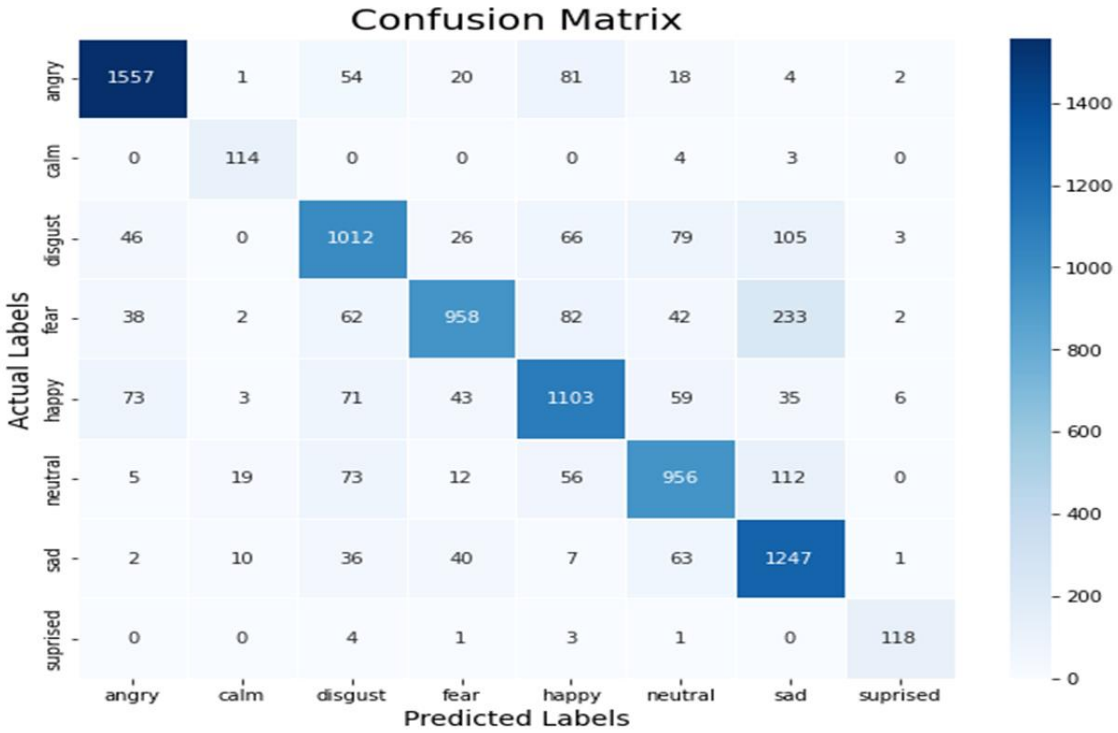


Fig. 5.4: Confusion Matrix (Actual vs. Predicted Labels)

5.5 Characteristics of Confusion Matrix

	precision	recall	f1-score	support
angry	0.90	0.90	0.90	1737
calm	0.77	0.94	0.84	121
disgust	0.77	0.76	0.76	1337
fear	0.87	0.68	0.76	1419
happy	0.79	0.79	0.79	1393
neutral	0.78	0.78	0.78	1233
sad	0.72	0.89	0.79	1406
suprised	0.89	0.93	0.91	127
accuracy			0.81	8773
macro avg	0.81	0.83	0.82	8773
weighted avg	0.81	0.81	0.80	8773

Fig. 5.5: Characteristics/Metrics of Confusion Matrix

REFERENCES

[1] G.E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neutral Computation, vol.18, no.7, pp. 1527-1554, 2006.

[2] M.El Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases, "Pattern Recognition, vol.44, no.3, pp. 572-587,2011.

[3] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[4] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[5] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.

[6] S.S.Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.

[7] H.Hu, M.-X.Xu, and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," in Proceedings of IEEE ICASSP 2007, vol.4. IEEE, 2007, pp. IV-413.