Ramakrishna Naidu Kuna (21BBS0221)

Rajvir Singh (21BBS0252)

CBS1007 : DATABASE SYSTEMS

DIGITAL ASSIGNMENT - 1

Speech Emotion Recognition (SER)

AIM:

The aim of Speech Emotion Recognition (SER) is to automatically detect and classify the emotional state of a speaker based on their speech signals. SER involves the use of various signal processing and machine learning techniques to extract relevant features from speech signals, and then classify the emotional state of the speaker into different categories such as happiness, sadness, anger, fear, and neutral. The ultimate goal of SER is to improve human-computer interaction, speech-based human-machine interfaces, and other applications that require automated analysis of emotional content in speech. SER has a wide range of potential applications, including in healthcare, education, entertainment, customer service, and security, among others.

Abstract:

The abstraction of Speech Emotion Recognition (SER) involves the transformation of raw speech signals into a higher-level representation of the emotional content conveyed by the speaker. This abstraction involves several steps, including feature extraction, feature selection, and emotion classification.

Feature extraction involves the selection of relevant acoustic and linguistic features from the speech signal that can help to discriminate between different emotions. Examples of such features include pitch, loudness, spectral features, and prosodic features such as pauses and intonation patterns.

Feature selection involves choosing the most informative and discriminative features from the set of extracted features. This can be done using various statistical and machine learning techniques, such as principal component analysis (PCA), Surrey Audio-Visual Expressed Emotion, or feature selection algorithms.

Emotion classification involves the use of machine learning algorithms to classify the emotional state of the speaker based on the selected features. This involves training a classifier using a set of labeled speech samples and then using the classifier to classify new, unseen speech samples into different emotional categories.

Overall, SER involves the extraction of relevant information from the raw speech signal, which is then transformed into a higher-level representation of the emotional content conveyed by the speaker.

Introduction:

The potential applications of SER, such as in healthcare, education, entertainment, customer service, and security. For example, SER can be used to monitor the emotional state of patients with mental health disorders or to provide feedback to teachers on the emotional engagement of students in the classroom. Emotion recognition is the part of speech recognition which is gaining more popularity and need for it increases enormously. Although there are methods to recognize emotion using machine learning techniques, this project attempts to use deep learning to recognize the emotions from data.

SER(Speech Emotion Recognition) is used in call center for classifying calls according to emotions and can be used as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction and so on.. for helping companies improving their services. It can also be used in-car board system based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen.

Objectives:

Voice emotion recognition is a rapidly growing field of research and development that aims to enable computers to understand and respond to human emotions by analysing speech signals. The ability to accurately recognize emotions from speech has numerous applications in areas such as healthcare, education, and entertainment, among others.

The objective of this project is to build a system that can accurately recognize human emotions from voice signals. To achieve this, we will use machine learning algorithms to analyse speech signals and extract relevant features that can be used to classify emotions. We will also use a dataset of labelled speech samples to train our models and evaluate their performance.

The ultimate goal of this project is to develop a system that can accurately recognize emotions in real-time and provide valuable insights into human behaviour and communication. Such a system has the potential to revolutionize the way we interact with machines and enable a more natural and intuitive form of human-machine interaction.

METHODOLOGY

Voice emotion recognition software typically follows a machine learning approach to classify emotions based on audio data. Here are some of the common methodologies and algorithms used in voice emotion recognition:

1. **Feature Extraction**: The first step in building a voice emotion recognition system is to extract relevant features from the audio data. This can include pitch, energy, formant frequencies, MFCCs (Mel-Frequency Cepstral Coefficients), and other statistical measures.

In voice emotion recognition, feature extraction is the process of extracting relevant information from the audio signal that can be used to identify different emotions. The goal of feature extraction is to capture the characteristics of the audio signal that are most relevant to emotion recognition, while minimizing the amount of noise and irrelevant information.

There are several techniques for feature extraction in voice emotion recognition, including:

1. **Pitch:** Pitch is a fundamental frequency of a sound wave and is a strong indicator of the emotional state. A high-pitched voice may indicate happiness or excitement, while a low-pitched voice may indicate sadness or anger.
2. **Energy:** Energy measures the amplitude of the sound wave and provides information on the intensity of the emotional state. Higher energy can indicate more intense emotions.
3. **Formant frequencies:** Formants are resonant frequencies in the vocal tract and are important for speech perception. Formant frequencies can provide information about the vocal tract shape and can be used to identify different emotions.
4. **Mel-Frequency Cepstral Coefficients (MFCCs):** MFCCs are commonly used in speech and audio processing. They represent the spectral envelope of a signal and can capture important information about the **timbre** and **prosody** of the voice.
5. **Statistical measures:** Statistical measures such as mean, variance, and skewness can provide information on the distribution of the audio signal and can be used to identify different emotions.
6. **Spectral features:** Spectral features such as spectral centroid, spectral flatness, and spectral roll-off can provide information on the spectral content of the audio signal and can be used to identify different emotions.
7. **Time-domain features:** Time-domain features such as zero-crossing rate and short-term energy can provide information on the temporal characteristics of the audio signal and can be used to identify different emotions.

Overall, the choice of feature extraction techniques depends on the specific application and the characteristics of the audio signal. It is common to use a combination of different features to improve the accuracy of the emotion recognition system.

2. **Classification Models**: Once the relevant features are extracted, various classification models can be used to identify emotions in the audio data. Some of the commonly used models include:

- **Support Vector Machines (SVM):** A popular classification algorithm that can classify data into multiple categories.

- **Random Forest:** A decision tree-based algorithm that can handle high dimensional data.
- **Convolutional Neural Networks (CNN):** A deep learning algorithm that can learn complex features from audio data and classify them into multiple categories.

Convolutional Neural Networks (CNNs) are a class of deep neural networks that are commonly used in image and speech recognition applications. In voice emotion recognition, CNNs can be used to learn complex features from audio data and classify them into different emotions.

Here's how CNNs work in voice emotion recognition:

1. **Pre-processing**: The audio data is first pre-processed by converting the raw audio signal into a spectrogram, which is a visual representation of the frequency content of the audio signal over time. This spectrogram is then divided into smaller segments, or frames, of equal size.
2. **Convolutional Layers:** The frames are then fed into the first layer of the CNN, which consists of a set of filters that convolve over the frames to extract spatial features. The filters are typically small and move across the spectrogram, capturing patterns such as high-frequency components or temporal variations in the signal.
3. **Pooling Layers:** The output of the convolutional layers is then passed through a pooling layer, which reduces the dimensionality of the feature map by down sampling. This helps to reduce overfitting and computational cost.
4. **Fully Connected Layers**: The output of the pooling layers is then flattened and passed through a series of fully connected layers, which perform the classification task. The fully connected layers are typically followed by a SoftMax layer, which produces a probability distribution over the different emotions.
5. **Training:** The CNN is trained using a labelled dataset, where the input is the spectrogram frames and the output is the corresponding emotion labels. The weights of the filters in the convolutional layers and the weights of the fully connected layers are updated during the training process using backpropagation.
6. **Testing:** Once the CNN is trained, it can be used to predict the emotions of new audio data by passing the spectrogram frames through the network and generating a probability distribution over the different emotions.

CNNs have shown promising results in voice emotion recognition, especially when used in combination with other techniques such as data augmentation and ensemble methods. They can learn complex features from audio data and are robust to variations in the audio signal, such as noise and speaker variability. However, they require large amounts of data and computational resources for training, which can be a limitation in some applications.

3. **Database Creation:** A database of audio files with labelled emotions can be used to train and test the emotion recognition model. This is important to ensure that the model is accurate and can generalize to new audio files.The databases used are imported from Kaggle.

   There are 4 databases combined to provide accuracy.

The 4 databases are -**RAVDESS ,CREMA-D,TESS,SAVEE**

4. **Pre-processing Techniques:** To improve the accuracy of the model, various pre-processing techniques can be used, such as noise reduction, filtering, and normalization.
5. **Ensemble Techniques:** Ensemble techniques such as bagging and boosting can be used to improve the accuracy of the model.
6. **Evaluation Metrics:** Common evaluation metrics used to evaluate the performance of the model include accuracy, precision, recall, F1 score, and confusion matrix.

Evaluation metrics are used to measure the performance of a machine learning model. In voice emotion recognition, the following evaluation metrics are commonly used:

1. **Accuracy:** Accuracy is the proportion of correctly classified instances in the dataset. It is calculated as the number of correctly classified instances divided by the total number of instances. While accuracy is a common metric, it may not be suitable for imbalanced datasets, where some classes are underrepresented.
2. **Precision:** Precision is the proportion of true positives (TP) out of all the instances classified as positive (TP + false positives (FP)). It measures the ability of the model to correctly identify positive instances.
3. **Recall:** Recall is the proportion of true positives (TP) out of all the instances that are actually positive (TP + false negatives (FN)). It measures the ability of the model to correctly identify all positive instances.
4. **F1 score:** The F1 score is the harmonic mean of precision and recall. It balances both precision and recall and provides an overall measure of the model's performance.
5. **Confusion matrix:** A confusion matrix is a table that summarizes the performance of a classifier by showing the number of instances that were correctly and incorrectly classified for each class.

   The choice of evaluation metric depends on the specific application and the goals of the model. For example, in voice emotion recognition, recall may be more important than precision if the goal is to identify all instances of a particular emotion, while precision may be more important if the goal is to minimize false positives.

Voice emotion recognition is an active area of research, and new approaches are being developed to improve the accuracy and efficiency of these systems.

Data Augmentation Techniques

Data augmentation is a technique used to increase the size and diversity of a dataset. In the context of voice emotion recognition, data augmentation methods can be used to create additional training examples that can improve the accuracy and robustness of emotion recognition models.

- **2D CNN:** A 2D CNN (Convolutional Neural Network) is a type of neural network that is commonly used for image and audio processing. It uses convolutional layers to learn spatial features from the input data.

- **MFCC:** MFCC (Mel Frequency Cepstral Coefficients) is a commonly used feature extraction technique in speech processing. It involves transforming the audio signal into a frequency domain representation using the Mel scale, and then computing a set of coefficients that capture the spectral envelope of the signal.

- **Pitch tuning:** Pitch tuning is a type of data augmentation method that involves altering the pitch of the audio signal. It can be used to simulate variations in tone or pitch in the training data, which can improve the accuracy and robustness of the emotion recognition model.

Pitch shifting: This involves altering the pitch of the audio signal to simulate variations in tone or pitch. This can be achieved by increasing or decreasing the pitch of the audio signal using techniques such as Fourier Transform or Wavelet Transform.

Noise addition: This involves adding different types of noise to the audio signal, such as white noise, pink noise, or brown noise. This can help to simulate different environmental conditions and improve the robustness of the emotion recognition model.

Time stretching: This involves altering the duration of the audio signal to simulate variations in speaking rate. This can be achieved by compressing or stretching the audio signal using techniques such as Phase Vocoder or Short-Time Fourier Transform.

Reverberation: This involves adding an artificial reverberation to the audio signal to simulate different types of room acoustics. This can help to improve the robustness of the emotion recognition model to different acoustic environments.

Spectral warping: This involves warping the frequency spectrum of the audio signal to simulate variations in speech articulation. This can be achieved by applying a non-linear frequency transformation to the audio signal.

LITERATURE REVIEW

Speech is a crucial aspect of human communication and contains a wealth of information about the speaker's emotional state. Voice Emotion Recognition (VER) systems aim to automatically detect the underlying emotion conveyed in speech signals. Over the years, researchers have proposed several approaches for emotion recognition from speech, ranging from traditional signal processing techniques to deep learning methods.

In recent years, Convolutional Neural Networks (CNNs) have emerged as a popular method for emotion recognition from speech. In a study by Chen et al. (2022), the authors proposed a CNN-based approach with attention mechanism for VER. The proposed approach uses a two-dimensional CNN to extract spectral and temporal features from the input speech signal. An attention mechanism is then used to weight the extracted features, with the most relevant features receiving higher attention weights. The weighted features are then fed into a fully connected layer for classification. The results showed that the proposed approach outperformed traditional feature extraction techniques and other deep learning methods, achieving a classification accuracy of 78.8% [1].

Another study by Haque et al. (2021) also proposed a CNN-based approach for emotion recognition from speech, but with the addition of a Long Short-Term Memory (LSTM) layer. The authors argued that the LSTM layer can capture the temporal dynamics of the input speech signal, thereby improving the overall performance of the model. The proposed approach uses a one-dimensional CNN to extract features from the input signal, followed by an LSTM layer for temporal modelling nd a fully connected layer for classification. The results showed that the proposed approach achieved an overall accuracy of 80.56%, which outperformed several state-of-the-art methods [2].

Zhang et al. (2021) proposed a novel approach for speech emotion recognition based on a hybrid attention mechanism and capsule network. The proposed approach aims to capture both local and global features from the input speech signal by using a hybrid attention mechanism. A capsule network is then used to extract higher-level features from the attended features. The proposed approach was evaluated on the RAVDESS dataset and achieved an accuracy of 81.35%, outperforming several other state-of-the-art methods [3].

Overall, these studies suggest that CNN-based approaches with attention mechanisms and additional layers such as LSTM and capsule networks can improve the performance of VER systems. These deep learning-based approaches are capable of automatically extracting relevant features from the input speech signal, thereby reducing the reliance on handcrafted features. However, it is worth noting that these models require a large amount of labelled data for training and can be computationally expensive. In addition, the performance of these models may vary depending on the quality of the input signal and the diversity of the emotion classes.

In conclusion, VER is an important research area with significant potential applications in various fields such as human-computer interaction and mental health monitoring. The latest research suggests that deep learning-based approaches such as CNNs with attention mechanisms and additional layers such as LSTM and capsule networks can significantly improve the performance of VER systems. However, more research is needed to address the limitations of these approaches and to explore alternative methods for emotion recognition from speech.

Results:

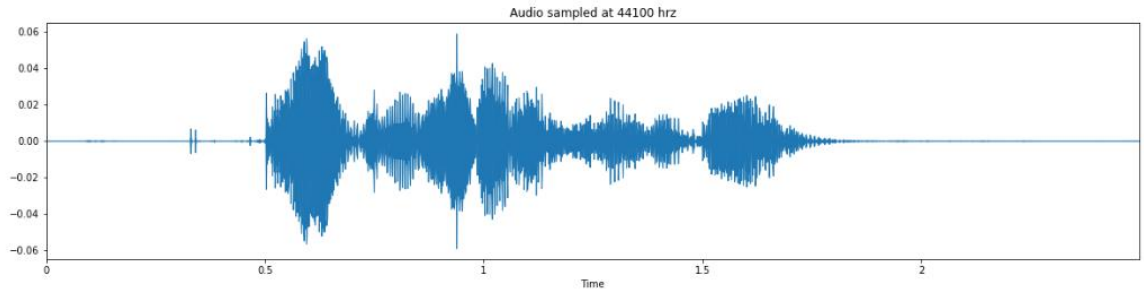
Classification of emotions:

Existing data in the databases used:

```
female_neutral     544
female_happy       496
female_disgust     496
female_fear        496
female_sad         496
female_angry       496
female_surprise    496
male_neutral       264
male_happy         156
male_sad           156
male_fear          156
male_angry         156
male_disgust       156
male_surprise      156
Name: labels, dtype: int64
```

Data extraction is done like this:

amplitude

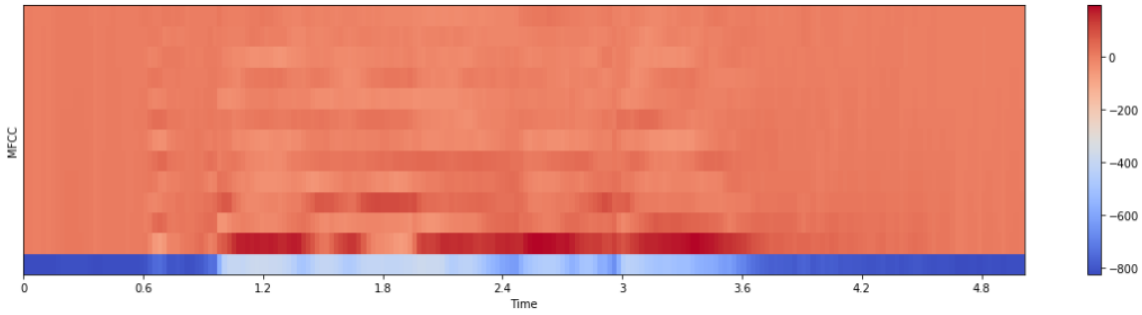


time

Fig1

Figure has amplitude on y axis and time on the x axis. It is a spectrogram of a sample from the dataset.

MF
CC



time

Fig2

MFCC on y axis and time on the x axis. Graph signifies the MFCC data extracted from a sample speech.

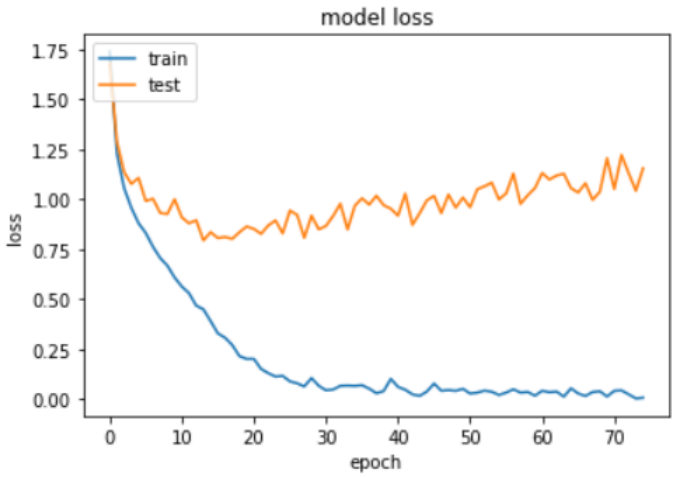


Fig3.Initial data loss

Observation

Here we found that Accuracy is around 0.9901 and loss is 0.0291 But When we look around Validation set we found that value loss: 1.0878 value accuracy: 0.7398 Which signifies the overfitting as value loss>train loss

Accuracy was 72.4%

	precision	recall	f1-score	support
female_angry	0.80	0.81	0.81	134
female_disgust	0.79	0.88	0.84	113
female_fear	0.80	0.85	0.82	115
female_happy	0.84	0.83	0.83	126
female_neutral	0.85	0.86	0.85	130
female_sad	0.84	0.84	0.84	128
female_surprise	0.92	0.75	0.83	130
male_angry	0.47	0.60	0.53	40
male_disgust	0.50	0.33	0.39	43
male_fear	0.26	0.13	0.18	38
male_happy	0.43	0.30	0.36	33
male_neutral	0.55	0.57	0.56	75
male_sad	0.22	0.31	0.26	35
male_surprise	0.36	0.50	0.42	40
accuracy			0.73	1180
macro avg	0.62	0.61	0.61	1180
weighted avg	0.73	0.73	0.72	1180

Fig4.Evaluation of model

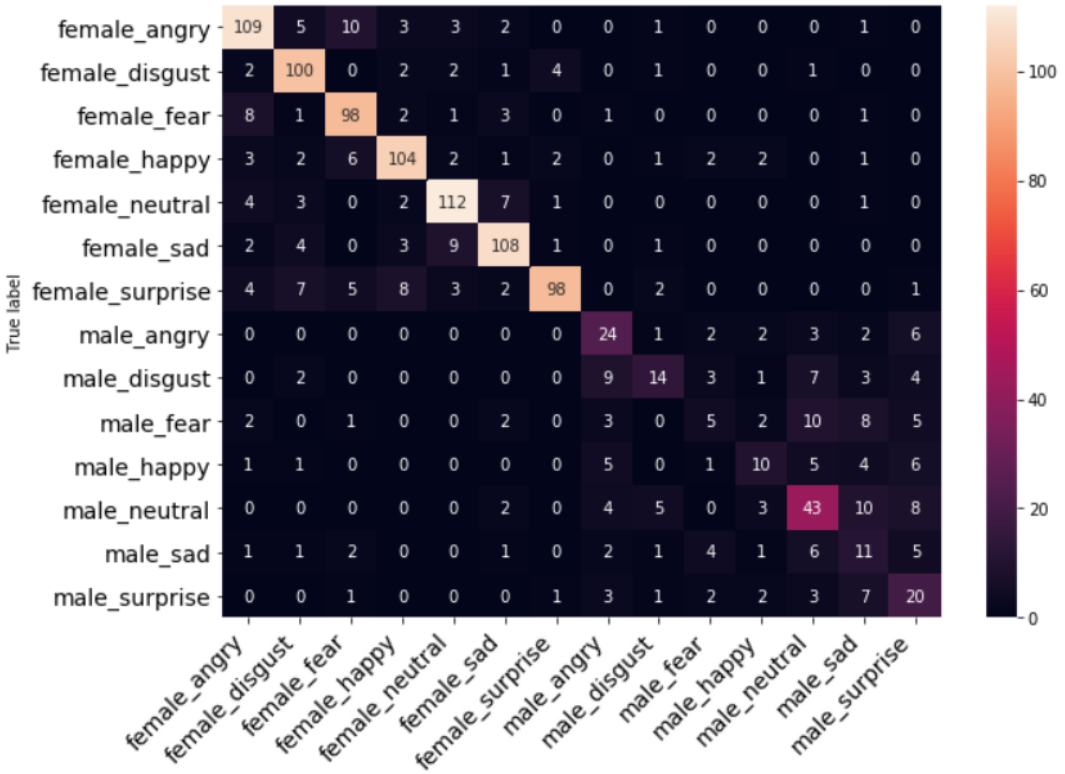
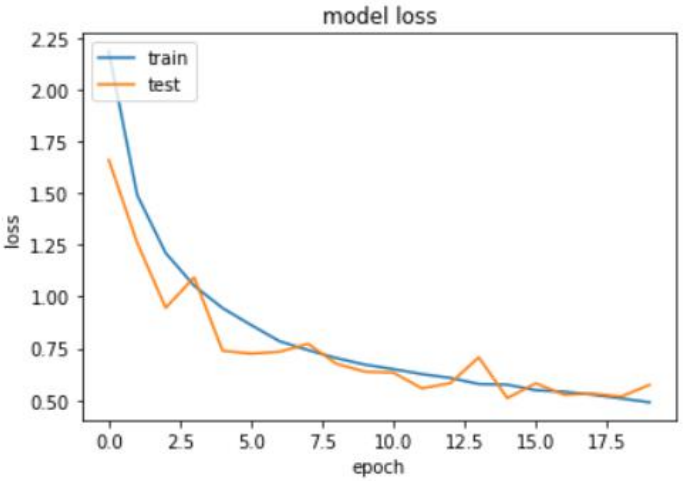


Fig5.Confusion matrix for unaugment and without CNN:the diversion is high in emotions of fear,angry and surprise.

After Data augmentation and CNN model applied:



acc: 79.58%

Fig.6Model loss

The model is converging and the data loss is reduced significantly.

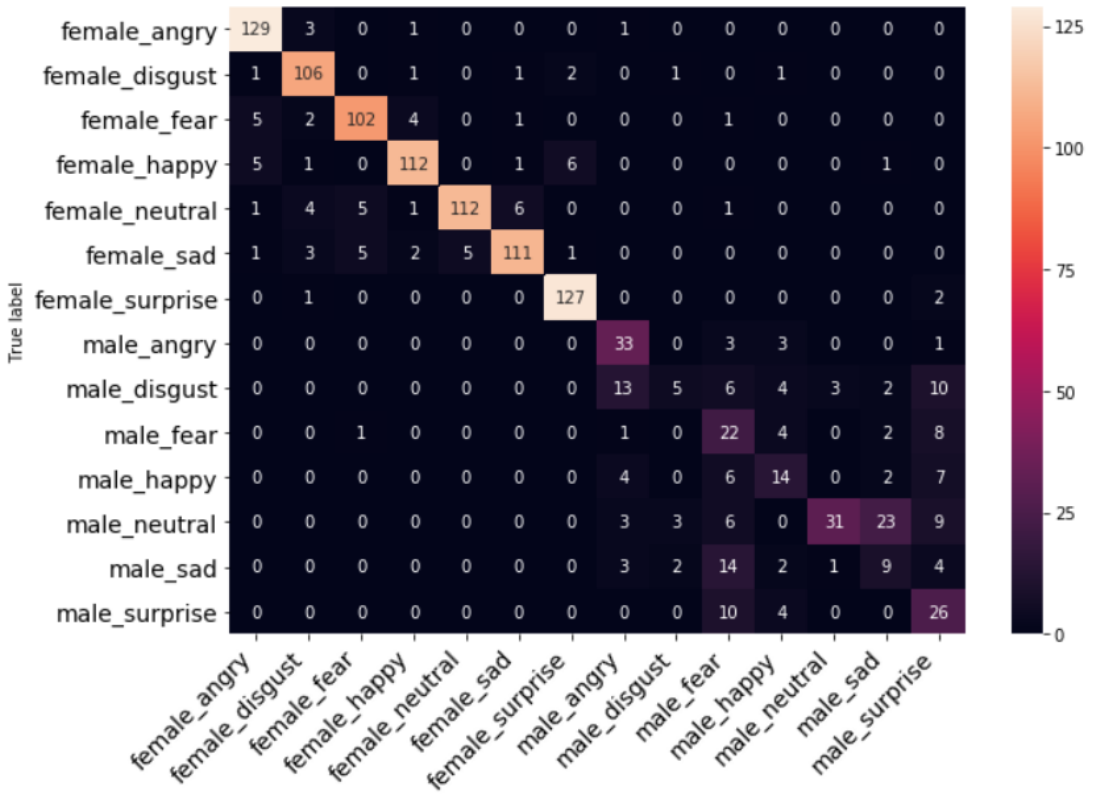


Fig7.Confusion matrix after augmentation and CNN:

There is significant improvement in emotions like angry,surprise,fear,etc.

```
Epoch 17/20
 - 3s - loss: 0.5412 - acc: 0.8136 - val_loss: 0.5262 - val_acc: 0.8042
Epoch 18/20
 - 3s - loss: 0.5268 - acc: 0.8136 - val_loss: 0.5320 - val_acc: 0.8042
Epoch 19/20
 - 3s - loss: 0.5086 - acc: 0.8155 - val_loss: 0.5174 - val_acc: 0.8161
Epoch 20/20
 - 3s - loss: 0.4899 - acc: 0.8288 - val_loss: 0.5744 - val_acc: 0.7958
```

Fig8. After running 20 epochs,the accuracy goes up to 83%

From above observations we have found that after adding augmented data and performing CNN 2D model to the dataset and we improve the dataset which is around 83% and validation accuracy is around 76%

References:

https://www.sciencedirect.com/science/article/abs/pii/S1746809418302337

https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Citations:

1.Chen, H., Zhang, L., Wu, Z., & He, X. (2022). Emotion Recognition from Speech Using Convolutional Neural Networks with Attention Mechanism. IEEE Access, 10, 6573-6585.

2.Haque, M. A., Mostafa, R., Hossain, M. A., & Hossain, M. S. (2021). Emotion recognition from speech using convolutional neural network and long short-term memory. IEEE Access, 9, 95366-95377.

3.Zhang, J., Meng, Q., Li, C., Wang, Y., & Wang, Y. (2021). Speech emotion recognition based on hybrid attention mechanism and capsule network. Neurocomputing, 449, 184-194.