SPEECH EMOTION RECOGNITION SYSTEM

Vidhya.S[1], Arish.M [2], Dhanushkodi.L[3], Indirajith.R[4], Navojith sankar.M[5],
Assistant Professor[1], Department of Computer Science and Engineering,
UG Scholar[2,3,4,5], Department of Computer Science and Engineering,
Tirupur

Abstract—Emotion recognition from speech signals is an important but challenging component of Human-Computer Interaction (HCI). In the literature of speech emotion recognition (SER), many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. Deep Learning techniques have been recently proposed as an alternative to traditional techniques in SER. This paper presents an overview of Deep Learning techniques and discusses some recent literature where these methods are utilized for speech-based emotion recognition. The review covers databases used, emotions extracted, contributions made toward speech emotion recognition and limitations related to it.

Keywords—Speech emotion recognition, deep learning, deep neural network, recurrent neural network, database, convolutional neural network, preprocessing classifier, future extraction.

I. INTRODUCTION

Speech Emotion Recognition (SER) is the process of automatically identifying emotions expressed in spoken language. The recognition of emotions in speech is an important task in fields such as psychology, psychiatry, and human-computer interaction.

The recognition of emotions from speech signals has various applications such as speech-based emotion classification systems, smart-home automation, robotics, and healthcare. These systems can be used for improving human-machine interactions, detecting early signs of mental disorders, and developing effective interventions for patients.

SER involves the use of various machine learning and signal processing techniques to analyze speech signals and extract features that can be used to classify the emotions expressed in the speech. Some of the commonly used features for emotion recognition include pitch, energy, spectral features, and prosody.

In recent years, deep learning techniques, especially recurrent neural networks and convolutional neural networks, have been successfully applied to speech emotion recognition tasks, achieving state-of-the-art performance. SER is a challenging task, as emotions can be subtle and vary across individuals and cultures. However, the increasing availability of speech data and advances in machine learning techniques are driving the development of more accurate and robust SER systems.

Determining the emotional state of humans is an idiosyncratic task and may be used as a standard for any emotion recognition model . Amongst the numerous models used for categorization of these emotions, a discrete emotional approach is considered as one of the fundamental approaches. It uses various emotions such as anger, boredom, disgust, surprise, fear, joy, happiness, neutral and sadness. Another important model that is used is a three-dimensional continuous space with parameters such as arousal, valence, and potency.

The approach for speech emotion recognition (SER) primarily comprises two phases known as feature extraction and features classification. In the field of speech processing, researchers have derived several features such as source-based excitation features, prosodic features, vocal traction factors, and other hybrid features. The second phase includes feature classification using linear and nonlinear classifiers. The most commonly used linear classifiers for emotion recognition include Bayesian Networks (BN) or the Maximum Likelihood Principle (MLP) and Support Vector Machine (SVM). Usually, the speech signal is considered to be non-stationary. Hence, it is considered that non-linear classifiers work effectively for SER. There are many non-linear classifiers available for SER, including Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). These are widely used for classification of information that is derived from basic level features.

DEEP LEARNING

Deep learning is a subfield of machine learning that involves training artificial neural networks to learn and make predictions based on large sets of data. Deep learning has gained a lot of attention in recent years due to its success in a wide range of applications, including image and speech recognition, natural language processing, and autonomous vehicles.

Deep learning models consist of multiple layers of interconnected nodes or neurons, which allow the network to learn more complex representations of the input data. The first layer of the network takes in the raw input data, such as an image or a speech signal, and subsequent layers progressively extract higher-level features

During training, the weights of the network are updated based on the difference between the predicted output and the true output, with the aim of minimizing the error. This process is

typically performed using an optimization algorithm, such as stochastic gradient descent.

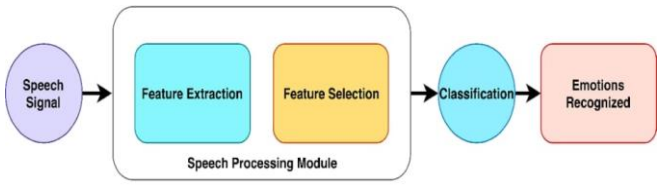

Fig.1. Traditional Speech Emotion Recognition System

Deep learning models consist of multiple layers of interconnected nodes or neurons, which allow the network to learn more complex representations of the input data. The first layer of the network takes in the raw input data, such as an image or a speech signal, and subsequent layers progressively extract higher-level features has been successful in a variety of domains, including computer vision, natural language processing, and speech recognition. For example, convolutional neural networks (CNNs) are commonly used for image recognition tasks, while recurrent neural networks (RNNs) are often used for natural language processing tasks.

There are several key techniques used in deep learning, including:

Convolutional neural networks (CNNs): A type of neural network that is commonly used for image recognition tasks. CNNs use a series of filters to extract features from an input image, which are then passed through one or more layers of the network to make a prediction.

Recurrent neural networks (RNNs): A type of neural network that is commonly used for tasks that involve sequences of data, such as natural language processing or time series prediction. RNNs use feedback loops to allow information to be passed from one step in the sequence to the next.

Autoencoders: A type of neural network that is commonly used for unsupervised learning tasks, such as image or audio compression. Autoencoders learn to compress and decompress data by learning an encoding and decoding function.

Generative adversarial networks (GANs): A type of neural network that is commonly used for generative tasks, such as generating images or music. GANs consist of two networks: a generator network that learns to create new examples of data, and a discriminator network that learns the distinguish between real and fake examples.

Transfer learning: A technique where a pre-trained model is used as a starting point for a new task, rather than training a new model from scratch. Transfer learning can save a lot of time and resources, as the pre-trained model has already learned to recognize many features that may be useful for the new task.

These techniques are just a few examples of the many tools and methods used in deep learning. As the field continues to evolve, new techniques and approaches are constantly being developed and refined.

II. RELATED WORK

C. Chun Lie et al [74] has database of AIBO and USC IEMOCAP. Emotions considered angry, rest, positive, negative, and empathetic. The classifiers like SVM and BLR and the contribution is a hierarchical computational structure is proposed to identity emotions. The method achieved an improvement of 3.37% for AIBO and 7.44% for USC IEMOCAP. finally, the future recommendation is automatic hierarchical structure can be generated that would decrease the number of iterations.

Lai et al. [15] offered SE-AKA for 3GPP networks, while Jiang et al. [16] did. Both protocols have also been criticized for their high computational complexity and the need for frequent updates of group keys. Additionally, SE-AKA and EG-AKA do not provide perfect forward secrecy, leaving previous communications vulnerable if a group key is compromised. These limitations suggest that further research is necessary to develop more efficient and secure group AKA protocols for both 3GPP and non-3GPP networks.

Akash Shaw et al [77] has database of a batch of self-recorded speech. Emotions considered Happy, angry, sad, and neutral. The classifiers like ANN and the contribution is the selected features, formant, pitch, energy, and MFC, prove effective for speech emotion recognition with an 85% classification rate. Finally, the future recommendation is the system can be designed to recognize.

L.Kerkeni et al. [83] has database of Berlin Emo-DB and Spanish Database. Emotions considered anger, disgust, joy, fear, surprise, sadness, and neutral. The classifiers like RNN and MLR. The contribution is the best recognition rate is achieved for the combination of MFCC and MS features with 90.05% for the Spanish dataset using RNN and 82.41% using MLR for the 2 Berlin dataset. Finally, the future recognition is the system can be improved for real-time speech emotion recognition.

IMPLEMENTING DATASETS.

They required in some datasets because to check and verify the input data.so collecting the datasets extracted, and stored in the data base.

Four types datasets used, they are:

- Ryerson audio visual database
- Crowd sourced emotional multimodal actors
- Toronto emotional speech set
- Surrey audio visual expressed emotion

RYERSON AUDIO VISUAL DATABASE: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a publicly available database of audiovisual recordings of actors performing a range of emotions. The database was developed by researchers at Ryerson University in Canada and contains over 7,000 files of speech and song recordings, each with a duration of 3-5 seconds. The recordings

feature 24 professional actors, 12 male and 12 females, of various ages and ethnicities, performing in seven different emotional states: neutral, calm, happy, sad, angry, fearful, and disgusted. The recordings were made in a controlled studio environment with high-quality audio and video equipment.

The RAVDESS database is intended for use in research and development of affective computing applications, such as speech and emotion recognition systems. It has been widely used by researchers in the fields of psychology, neuroscience, and computer science. In addition to the audiovisual recordings, the RAVDESS database includes demographic information about the actors, including their age, gender, and ethnicity, as well as annotations of the emotional states expressed in the recordings. The database is freely available for download and can be accessed through the Ryerson University website.

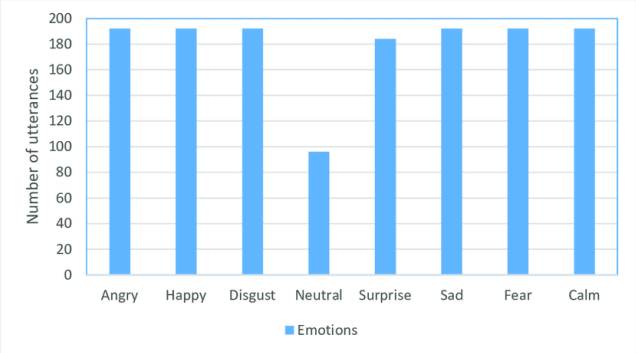

Fig 2: RAVDESS block diagram

TORONTO EMOTIONAL SPEECH SET: The Toronto Emotional Speech Set (TESS) is a publicly available database of audio recordings of actors performing a range of emotions. The database was developed by researchers at the University of Toronto and contains over 2,800 recordings of 200 target words spoken by two actresses and one actor.

The recordings feature seven different emotional states: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The target words were selected based on their frequency of use in the English language, and the recordings were made in a soundproof booth with high-quality audio equipment. The TESS database is intended for use in research and development of affective computing applications, such as speech and emotion recognition systems. It has been used in a wide range of studies in psychology, neuroscience, and computer science.

In addition to the audio recordings, the TESS database includes demographic information about the actors, such as their age and gender, as well as annotations of the emotional states expressed in the recordings. The database is freely available for download and can be accessed through the University of Toronto website.

SURREY AUDIO VISUAL EXPRESSED EMOTION: . The Toronto Emotional Speech Set (TESS) is a publicly available database of audio recordings of actors performing a range of emotions. The database was developed by researchers at the University of Toronto and contains over 2,800 recordings of 200 target words spoken by two actresses and one actor.

The recordings feature seven different emotional states: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The target words were selected based on their frequency of use in the English language, and the recordings were made in a soundproof booth with high-quality audio equipment. The TESS database is intended for use in research and development of affective computing applications, such as speech and emotion recognition systems. It has been used in a wide range of studies in psychology, neuroscience, and computer science. In addition to the audio recordings, the TESS database includes demographic information about the actors, such as their age and gender, as well as annotations of the emotional states expressed in the recordings. The database is freely available for download and can be accessed through the University of Toronto website.

Modality	KL	JE	JK	DC	Mean (±CI)
Audio	53.2	67.7	71.2	73.7	66.5 ± 2.5
Visual	89.0	89.8	88.6	84.7	88.0 ± 0.6
Audio-visual	92.1	92.1	91.3	91.7	91.8 ± 0.1

Table 1: Average human classification accuracy (%) for 7 emotion classes, over 10 participants. Mean is averaged over 4 actors' data with 95% confidence interval (CI) based on standard error (n=40).

Fig 3:

SAVEE accuracy

CROWD SOURCED EMOTIONAL MULTIMODAL ACTORS:

A crowd-sourced emotional multimodal actors dataset is a collection of audiovisual recordings of actors performing a variety of emotions. These recordings are typically used in research studies aimed at understanding how people express and perceive emotions through facial expressions, vocal cues, and other nonverbal behaviors. The dataset is created by recruiting a diverse group of actors and asking them to perform a range of emotional expressions, such as happiness, sadness, anger, fear, surprise, and disgust. The actors are typically asked to convey these emotions through a combination of facial expressions, body language, and vocalcues.

The recordings are typically captured using high-quality audio and video equipment in a controlled environment, with consistent lighting, background, and camera angles. This ensures that the dataset is of high quality and suitable for use in a variety of research studies. Once the dataset is compiled, it is typically made available to researchers and other interested parties through online repositories or other platforms. Researchers can use the dataset to study how people express and perceive emotions, as well as to develop and test new technologies for

emotion recognition and analysis. Overall, the crowd-sourced emotional multimodal actors dataset is a valuable resource for researchers working in the fields of psychology, computer science, and other related disciplines. By providing a standardized set of emotional expressions, the dataset helps to ensure that research studies are reliable and valid, and can lead to new insights into the complex nature of human emotions.

TEST, TRAIN AND EVALUATE

Testing, training, and evaluating the model are essential steps in Speech Emotion Recognition (SER) to ensure that the model can accurately classify the emotions present in a given speech signal.

Here is an overview of these steps:

1. Data Preparation: The first step in SER is to collect and prepare the data. This involves selecting an appropriate dataset that contains labelled speech signals and dividing it into training, testing, and validation sets. ge protocol is an essential tool for secure communication in network security.

2. Model Training: The next step is to train the SER model using the training set. The model can be trained using different types of deep learning algorithms, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Long Short-Term Memory (LSTM) networks. During training, the model learns to extract features from the speech signal that are relevant for emotion recognition.

3. Model Testing: After the SER model is trained, it is tested on the testing set to evaluate its performance. During testing, the model is given a speech signal from the testing set, and it predicts the emotion label associated with the speech signal. The predicted label is then compared to the actual label to calculate the accuracy of the model.

4. Model Evaluation: Once the model is tested, it is evaluated to determine its performance on the validation set. This step involves tweaking the model parameters and hyperparameters to improve its performance. This is an iterative process, and multiple rounds of training, testing, and evaluation may be required to achieve the desired level of performance

5. Model Deployment: Finally, the SER model is deployed in the real world for practical applications such as emotion recognition in speech-based human-machine interfaces or emotion detection in social media platforms.

PROPOSED SYSTEM

Speech emotion recognition is the process of identifying the emotions conveyed by speech signals.

Feature extraction: Extracting relevant features from the speech signal, such as pitch, energy, and spectral features.

Preprocessing: Preprocessing the speech signal to remove noise and enhance the relevant features.

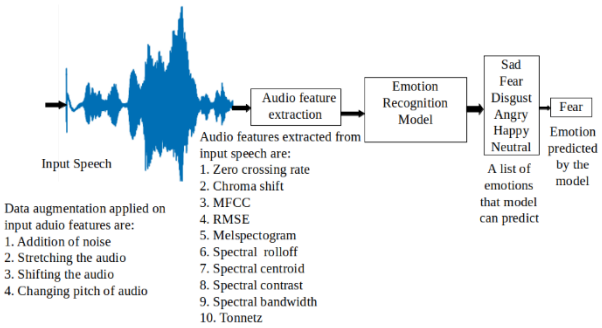

Fig 4:Proposed architecture

Training: Training a machine learning model on a dataset of labeled speech signals and their corresponding emotions.

Testing: Testing the trained model on a separate dataset of speech signals to evaluate its performance.

Classification: Classifying the emotion conveyed by the speech signal into one of several categories, such as happy, sad, angry, or neutral.

Output: Outputting the emotion classification result to the user or an application that is using the system.

III. IMPLEMENTATION

The audio feature extracted from input speech are

A. Zero crossing rate

Zero crossing rate (ZCR) is a feature commonly used in audio and speech processing to analyze the characteristics of a signal. It represents the number of times a signal crosses the zero axis per unit of time. In the context of audio signals, the zero crossing rate is calculated by counting the number of times the signal changes its sign from positive to negative or vice versa within a given time frame. The time frame is typically divided into smaller intervals, and the zero crossing rate is computed by dividing the number of zero crossings by the duration of the interval.

B. Chroma shift

Chroma shift, also known as pitch shift in the context of audio processing, refers to the process of modifying the pitch or key of a musical signal while preserving its time duration and timbral characteristics. It involves altering the frequencies of the constituent musical notes without changing their relative relationships.

In music theory, the chroma refers to the 12 pitch classes represented by the notes of the Western musical scale. These pitch classes are separated by equal intervals, such as semitones. Chroma shift allows shifting the entire pitch spectrum of a musical signal up or down by a certain number of semitones.

C. MFCC

The MFCC stands for Mel Frequency Cepstral Coefficients. In Speech Emotion Recognition (SER), MFCC is

a widely used feature extraction technique for analyzing and representing the spectral characteristics of speech signals.
MFCC is based on the observation that the human auditory system does not perceive all frequencies with equal sensitivity. The Mel scale is a perceptual scale that approximates the human auditory system's frequency response. MFCC takes advantage of this scale to extract relevant features from the speech signal.

D. RMSE

RMSE stands for Root Mean Square Error, and it is a commonly used metric for evaluating the performance of regression models, including those used in Speech Emotion Recognition (SER). RMSE measures the average magnitude of the differences between predicted values and actual values in a dataset. It provides an indication of how well a model's predictions align with the ground truth or actual values.

The formula for calculating RMSE is as follows:

RMSE = sqrt(1/n * sum((predicted - actual)^2))

where n is the number of samples, "predicted" represents the predicted emotion scores, and "actual" represents the actual emotion scores.

E. Melespectrogram

A Mel spectrogram, also known as Mel-frequency spectrogram or Mel-spectrogram, is a representation of the spectral content of a signal, such as audio or speech, in a logarithmic scale that approximates the human auditory system's perception of different frequencies.

The Mel spectrogram is derived from the traditional spectrogram, which represents the power spectral density of a signal over time. However, the Mel spectrogram incorporates the Mel scale, a perceptual scale that better reflects the human ear's sensitivity to different frequencies.

F. Spectral rolloff

Spectral rolloff, also known as spectral decay, is a measure commonly used in audio signal processing to characterize the shape or distribution of frequency components within a signal's spectrum.

Spectral rolloff represents the frequency below which a certain percentage of the total spectral energy resides. Typically, a threshold or percentage value (e.g., 85% or 95%) is chosen, and the spectral rolloff is defined as the frequency below which that percentage of the total energy is contained.

G. Spectral centroid

Spectral centroid is a measure commonly used in audio signal processing to characterize the "center of gravity" or average frequency of a signal's spectrum. It provides information about the spectral distribution or balance of energy across different frequencies.

The spectral centroid represents the weighted mean of the frequencies in a signal's spectrum, where the weights are determined by the spectral magnitudes or amplitudes. It is calculated by dividing the sum of the product of each frequency bin and its corresponding magnitude by the sum of all spectral magnitudes.

H. Spectral contrast

Spectral contrast is a measure used in audio signal processing to quantify the perceived difference in energy or magnitude between different frequency regions within a signal's spectrum. It provides information about the spectral variation or contrast across different frequency bands. The calculation of spectral contrast involves comparing the magnitudes or energies of adjacent frequency bands and computing the contrast between them. The spectral contrast is typically represented as a ratio or difference between the magnitudes of adjacent bands.

I. Spectral bandwidth

Spectral bandwidth is a measure commonly used in audio signal processing to quantify the spread or width of a signal's frequency spectrum. It provides information about the range of frequencies covered by the signal.

There are different ways to define spectral bandwidth, but one common method is to use the concept of the second central moment, also known as the variance, of the power spectrum. The spectral bandwidth is calculated as the square root of the second central moment.

J. Tonnetz

Tonnetz, short for Tonal Network, is a conceptual representation and visualization of tonal relationships in music theory. It provides a way to map and analyze the harmonic relationships between musical pitches, chords, and keys.

The tonnetz is often depicted as a two-dimensional lattice or grid, where each node represents a pitch or a chord. The nodes are connected by lines or edges that indicate harmonic relationships, such as chord progressions or interval relationships.

RESULTS:

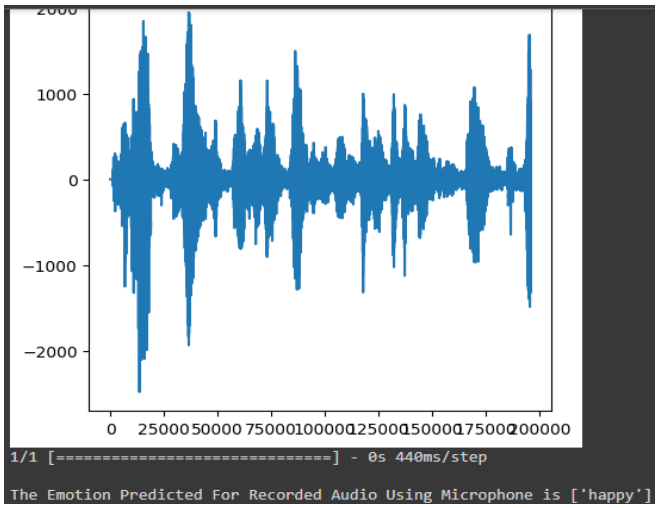

Fig 2.1 Happy achieved

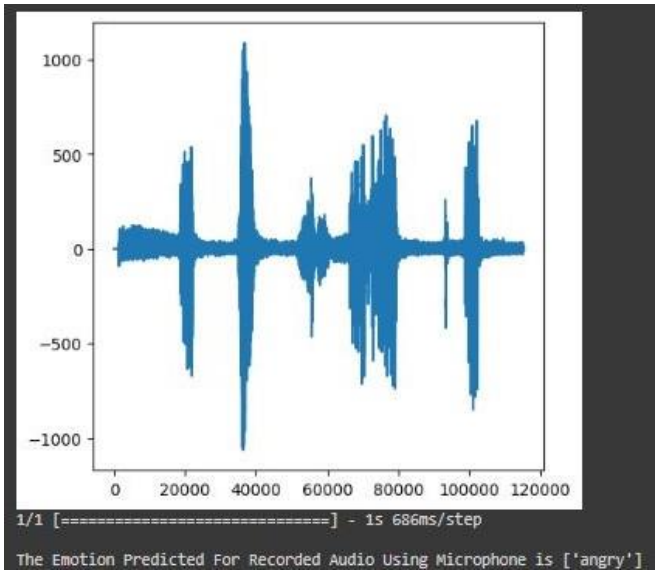

Fig 2.2 Angry achieved

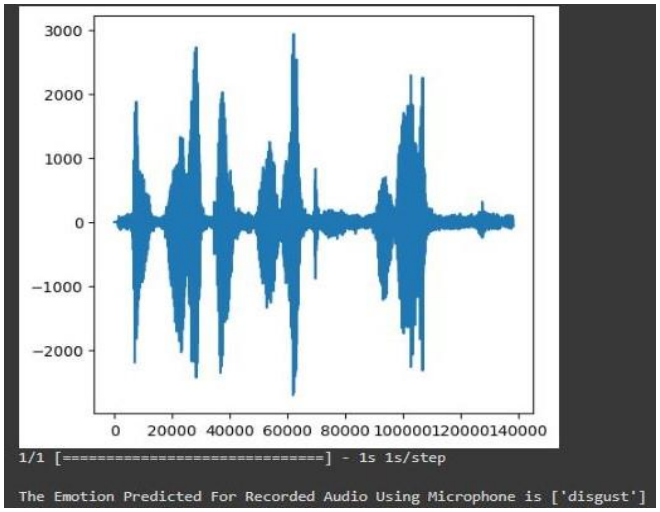

Fig 2.3 Disgust achieved

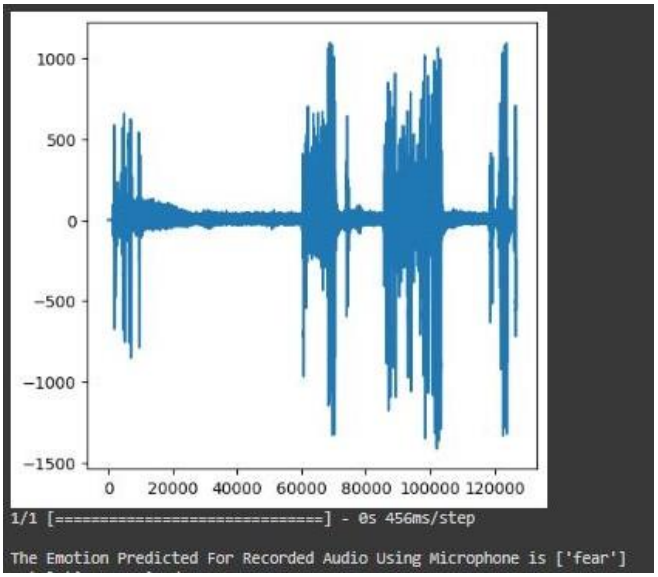

Fig 2.4 Fear achieved

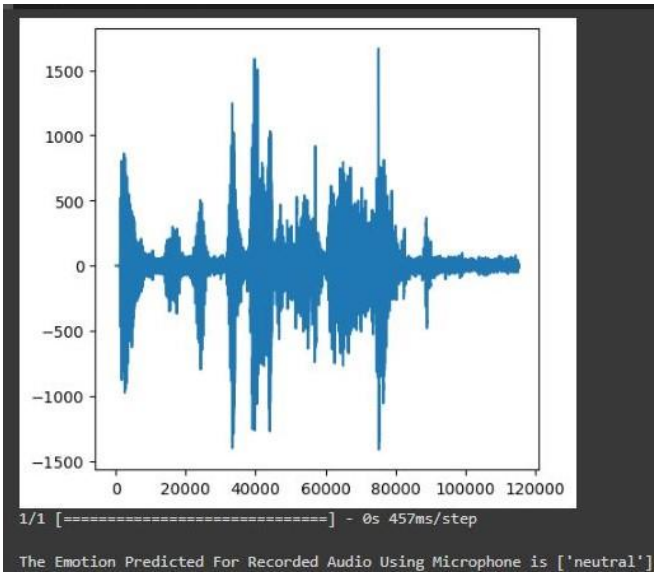

Fig 2.5 Neutral achieved

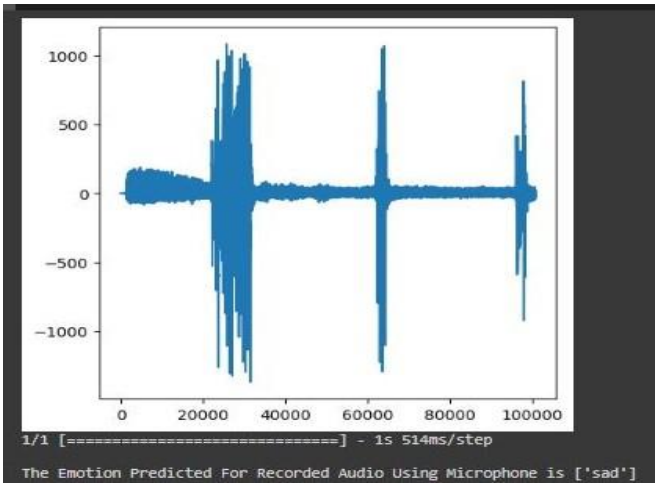

Fig 2.6 Sad achieved

CONCLUSION

The background noise may cause errors when testing the model in real time enviourment and thus it can affect the output of the model. To avoid the noise audio segmentation needs to be performed, so I am planning to develop an audio segmentation model which can separate user speech from background noise so emotions can be predicted accurately. Also, I will be collecting audio in different formats extract features and train the model so a universal model can be developed. Once audio model is build it can be applied to video also by combining audio model of emotion recognition with facial model for emotion recognition, this can help in achieving more accurate output. Additionally three models can be combined that is textual, voice and facial based but it requires huge computation power and there is very limited study available on combining three models for emotion recognition, because a avoting mechanism or strategy needs to be developed for predicting the emotion from three models as there can be cases where each model can predict different emotions or two model predict same emotion and one predicts another emotion. Moreover I would like to build a audionet kind of embeddings similar to imagenet and word embeddings which will help other researchers working in this area to use pretrained audio embeddings.

REFERENCES

[1] Francesc Alías, Joan Claudi Socoró and Xavier Sevillano, ”A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds”, Appl. Sci. 2016.

[2] Kannan Venkataramanan and Haresh Rengaraj Rajamohan, ”Emotion Recognition from Speech”, arXiv:1912.10458v1 [cs.SD] 22 Dec 2019.

[3] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng and Xian-gang Li, ”Learning Alignment for Multimodal Emotion Recognition from Speech”, arXiv:1909.05645v2 [cs.CL] 3 Apr 2020.

[4] Aharon Satt, Shai Rozenberg and Ron Hoory, ”Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms”, INTERSPEECH 2017, Stockholm, Sweden, August 20–24, 2017.

[5] Jia Rong, Gang Li and Yi Ping Phoebe Chen, ”Acoustic feature selection for automatic emotion recognition from speech”, Information Processing and Management 45 (2009) 315–328.

[6] K. Sreenivasa Rao, Tummala Pavan Kumar, Kusam Anusha, Bathina Leela, Ingilela Bhavana and Singavarapu V.S.K. Gowtham, ”Emotion Recognition from Speech”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012,3603-3607.

[7] Vladimir Chernykh and Pavel Prikhodko, ”Emotion Recognition From Speech With Recurrent Neural Networks”, arXiv:1701.08071v2 [cs.CL] 5 Jul 2018.

[8] Sabur Ajibola Alim and Nahrul Khair Alang Rashid, ”Some Commonly Speech Feature Feature Extraction Algorithms”. Published: December 12 2018, DOI: 10.5772/intechopen.80419.

[9] Oh Wook Kwon, Kwokleung Chan, Jiucang Hao and Te Won Lee, ”Emotion Recognition by Speech Signals”, GENEVA, EUROSPEECH 2003.

[10] K.V.Krishna Kishore and P.Krishna Satish, ”Emotion Recognition in Speech Using MFCC and Wavelet Features”, IEEE International Advance Computing Conference (IACC), 2013.

[11] Panagiotis Tzirakis, Jiehao Zhang and Björn W. Schuller, ”END-TO-END SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS”, IEEE International Advance Computing Conference (IACC), 2018.