Neural Network Based Approach For Recognition of Basic Emotions from Speech

Anusha A Naik[1], Anusha D[2], Bhagyashree M[3], Juhi Verma[4](Students SJBIT-ISE)
Guide-Mrs. Ashwini K(Assistant Professor, Dept. of ISE)
Department of Information Science and Engineering
SJB Institute of Technology[1,2,3,4], Visvesvaraya Technological University
Karnataka, India

***Abstract*— Emotional recognition from speech is one of the most researched topics nowadays in the field of signal processing and human machine interaction system.**

Unlike humans, machines do not have the ability to detect and express emotions. However, personal computer interaction can be improved with an automated emotional recognition system by reducing the need for human access. In this study paper, a speech recognition technique was introduced using the Convolutional Neural Network (CNN) with features of Mel Frequency Cepstral Coefficients (MFCC). Human-computer interaction can be improved with emotions recognition.

In this project, a method for speech emotion recognition is presented using Neural Network with Mel Frequency Cepstral Coefficients (MFCC) features and RAVDESS Dataset is used. From our project, we have found an average of 57.64% accuracy where 8 different emotions were classified.

***Index Terms*—Speech Emotion, MFCC, Convolutional Neural Network.**

I. INTRODUCTION

Emotional awareness is a term used to identify a person's emotions from a person's voice, frequency, and facial expressions. Machines cannot detect and express emotions as a person. But human interaction with computers can also be filtered through an automated emotional monitoring system, which can help reduce human interventions.Emotion recognition is the terminology for identifying human emotions from human voice intensity, pitch and frequency.Speech Emotion Recognition (SER) is the task of recognizing the emotional aspects of speech irrespective of the semantic contents. In this project, eight basic emotions (like happy, sad, neutral, angry, disgust,fearful,calm and surprise) are analyzed from emotional speech signals.

In a voice-based system, a computer agent is required to completely comprehend the human's speech percept in order to accurately pick up the commands given to it. This field of study is termed as Speech Processing and consists of three components: Speaker Identification, Speech Recognition,Speech Emotion Detection

The main focus of the speech recognition system is the signal processing unit where the relevant features are extracted from the available speech signal and the other is a system that detects sensory signal speech. The range of sensory perception in speech is small and limited. In this day and age, researchers are trying to find out which factors are the most important factor in influencing emotional recognition in speech. It is also unthinkable tofeature. The MFCC feature extracting method is used in this work. It is the best way to extract an element from speech. The mel-frequency cepstrum represents the short-term energy spectrum of sound. This cepstrum is based on a choose the best algorithm for emotional separation and which emotion will be shared together. The most popular methods are the Bayesian learning, the Convolutional Neural Network (CNN) [1], Linear Discriminant Analysis

Feature extracting is the process of gathering information from a set of discriminating samples. Pitch Detection Algorithm (PDA), Linear Predictive Coding (LPC), Mod- ulation Spectral Features (MSFs), Mel-Frequency Cepstral

Coefficients (MFCC), etc. are different ways to extract a direct cosine version of the log energy indicator on the indirect mel scale offrequencies. Mel-frequency cepsral coefficients are coefficient (LDA), vector support machine (SVM) [2] as a high-capacity LDA extension with the space feature and the Markov hidden model (HMM) [3] to capture temporary state changes. But now a day's neural networks play a more important role in sensitizing and displaying better energy than ever before. Deep Neural Network operates (DNN) [4] as Convolutional Neural Network (CNN) [5], Recurrent Neural Network (RNN) [6], Artificial Neural Network (ANN), and Convolutional Neural Network and Long Short-Term Memory (CNN- LSTM) are widely used these days. responsible for MFC joint formation. The Convolutional Neural Network (CNN) model is used for this function. Emotion recognition is the terminology for identifying human emotions from human voice intensity, pitch and frequency.Speech Emotion Recognition (SER) is the task of recognizing the emotional aspects of speech irrespective of the semantic contents. Here, eight basic emotions (like happy, sad, neutral, angry, disgust, fear, calm and surprise) are analyzed from emotional speech signals. Recognition of emotions from the speech is one of the most researched topics now a days in the field of signal processing and human machine interaction system.Machines can't perceive and show emotions as like as human so our aim was to build a model to recognize emotion from speech using the librosa and sklearn libraries and the ravdess dataset.

The primary objective of a project is to give out better accuracy and classify many emotions.

II. RELATED WORK

Automatic Speech Emotion is a novel research topic now a days which is mainly evolved in the Human Computer Interaction field. In past years several classification methods were used for speech emotion recognition.

In a research, Md. Sham-E-Ansari et al. [1]proposed a SER model using Neural Network with MFCC features and EmoDB was used for classification. Apart from this, Komal Raj vanshi et al. [8] proposed a model that used mathematical power, inclination, MFCC and entropy as features and neural network as a separator. Abdul Malik Badshah et al. [4] proposed DNN for speech recognition recognition in another paper. CNN was used in their model and achieved an 84.3% accuracy. Apart from this, Saikat Basu et al. [5] raised emotional awareness in speech using CNN with RNN structure and used the CNN-LSTM partitioning method.Their test accuracy was 80%.

Pavol Har'r et al. [9] described the Speech Recognition Method using DNN Architecture with convolutional, pooling, and fully integrated layers. They used the Voice Activity Detection (VAD) algorithm to remove silent segments. Sonali T. Saste et al. [10] used Discrete Wavelet Transform (DWT) and MFCC to extract the feature and SVM as a separator. Speech elements were extracted from two different figures MFCC and DWT. Panagiotis Tzirakis et al. [11] demonstrated end-to-end sensory recognition using deep neural networks in their research. Their model was comprised of CNN and LSTM. Wootaek Lim et al. [12] proposed speech recognition using Convolutional and Recurrent Neural Networks built with CNN, LSM and CNN Distributed Time using Short Time Fourier Transform (STFT). In another paper, Xiaomin Chen et al. [13] suggested sequential sequence to speech sensory modeling using the Connectionist Temporal Classification based Recurrent Neural Network (CTC-RNN).

Vladimir Chernykh et al. [14] used the one label method and the Connectionist Temporal Classification (CTC) method. The Bidirectional Long-Term Short-Term Memory (BLSTM) network was much more efficient than the simple LSTM network within this single label method and the BLSTM with CTC loss was twice as much training as possible compared to the entropy loss . Leila Kerkeni et al. [15] used the Multivariate Linear Regression (MLR) classifier, the SVM classifier, and the RNN classifier based on the Berlin and Spanish databases. The best result of the recognition level was 90.05%, obtained by combining the features of MFCC and MS.

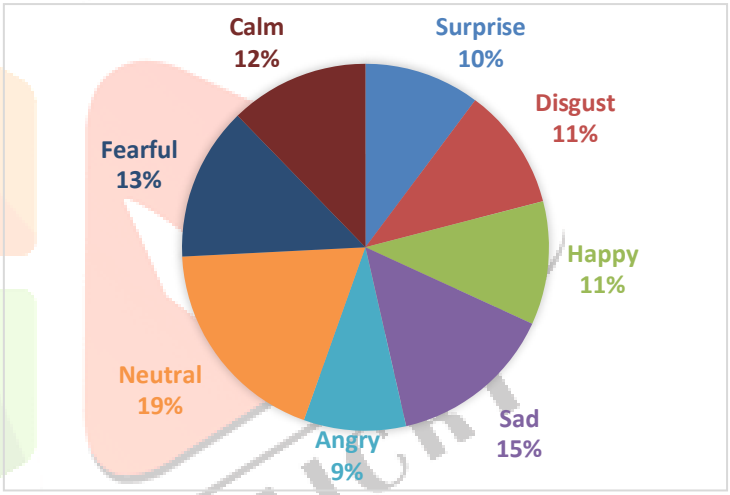
A. METHODOLOGY

B. Dataset and Data Virtualization

Here we have used The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Different speech samples will be collected of different humans with respect to different emotions.

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 includes 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

For the experiment, we have used a total of 1440 samples consist of 8 different emotions: Happy (120 recordings), Angry (90 recordings), Neutral (270 recordings), Sad (210 record ings), Calm (140 recordings), Surprise (150 recordings), Fearful (195 recordings) and Disgust (155 recordings). 80% of the samples are used for training the data and 60% we have used for testing the data.



C. Data Preprocessing

Digital-to-analog converters change the analog into binary digital signals. The expression and experience of human behavior are complex, multimodal and characterized by individual and contextual heterogeneity and variability. The preprocessing of speech consists of cleaning the speech signal from ambient and undesirable noises, detecting speech activity, and normalizing the length of the vocal tract. Signal processing consists in applying acoustic filters on original audio signals and splitting it into units.

The Discrete Fourier Transform is the most widely used transform in all areas of digital signal processing because it allows converting a sequence from the time domain to the frequency domain. DCT provides a convenient representation of the distribution of the frequency content of an audio signal. The use of this transform is crucial because the majority of audio features extracted to analyze speech emotion are defined in the frequency domain. Given a discrete-time signal xnxn, n=0n=0,……, N−1N−1, the Discrete FourierTransform can be defined as -

Xk=N−1∑n=0xke−i2πNknk=0,...,N−1Xk=∑n=0N−1xke−i 2πNknk=0,...,N−1.

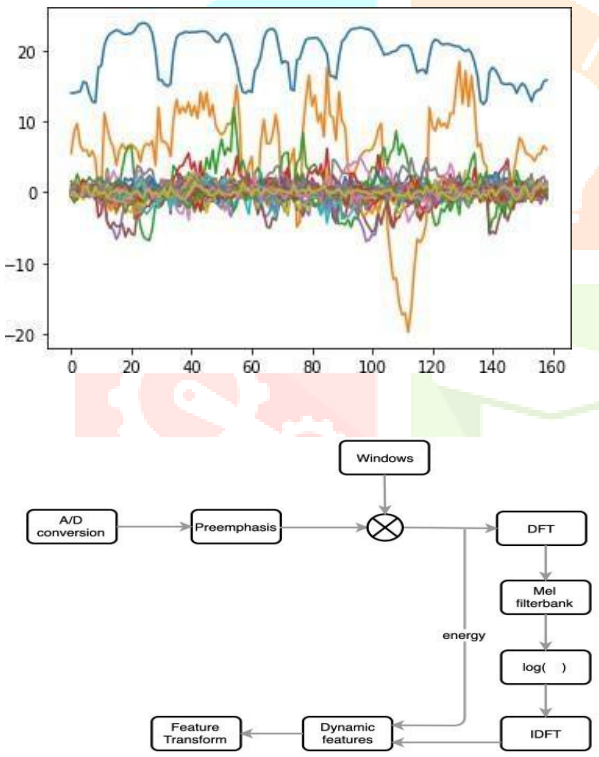
The Discrete Fourier Transform outputs sequence of N coefficient XkXk constituting the frequency domain representation of a signal. The inverse Discrete Fourier Transform takes Discrete Fourier coefficient and returns the original signal in the time-domain:

xn=1NN−1∑k=0xnei2πNknn=0,...,N−1

D. Feature Extraction

For feature extraction MFCC is used in this work.

Mel Frequency Cepstral Coefficient (MFCC) technique is used to recognize emotion of a speaker from their voice. The designed system was validated for Happy, Sad, Anger, Neutral, Calm, Surprise, Fearful and Disgust emotions and the efficiency was found to be about 57.64%.



MFCC for Feature ExtractionA/D Conversion:

In this step, we will convert our audio signal from analog to digital format with a sampling frequency of 8kHz or 16kHz.

Preemphasis:

Preemphasis increases the magnitude of energy in the higher frequency. When we look at the frequency domain of the audio signal for the voiced segments like vowels, it is observed that the energy at a higher frequency is much lesser than the energy in lower frequencies. Boosting the energy in higher frequencies will improve the phone detection accuracy thereby improving the performance of the model.

Windowing:

The MFCC technique aims to develop the features from the audio signal which can be used for detecting the phones in the speech. But in the given audio signal there will be many phones, so we will break the audio signal into different segments with each segment having 25ms width and with the signal at 10ms apart as shown in the below figure. On average a person speaks three words per second with 4 phones and each phone will have three states resulting in 36 states per second or 28ms per state which is close to our 25ms window. From each segment, we will extract 39 features. Moreover, while breaking the signal, if we directly chop it off at the edges of the signal, the sudden fall in amplitude at the edges will produce noise in the high-frequency domain. So instead of a rectangularwindow, we will use Hamming/Hanning windows to chop the signal which won't produce the noise in the high-frequency.

DFT(Discrete Fourier Transform):

We will convert the signal from the time domain to the frequency domain by applying the dft transform. For audiosignals, analyzing in the frequency domain is easier than in the time domain.

Mel-Filter Bank:

The way our ears will perceive the sound is different from how the machines will perceive the sound. Our ears have higher resolution at a lower frequency than at a higher frequency. So if we hear sound at 200 Hz and 300 Hz we can differentiate it easily when compared to the sounds at 1500 Hz and 1600 Hz even though both had a difference of 100 Hz between them. Whereas for the machine the resolution is the same at all the frequencies. It is noticed that modeling the human hearing property at the feature extraction stage will improve the performance of the model. So we will use the mel scale to map the actual frequency to the frequency that human beings will perceive. The formula for the mapping is given below.

Applying Log:

Humans are less sensitive to change in audio signal energy at higher energy compared to lower energy. Log function also has a similar property, at a low value of input x gradient oflog function will be higher but at high value of input gradient value is less. So we apply log to the output of Mel-filter to mimic the human hearing system.

$$mel(f) = 1127\ln(1+\frac{f}{700})$$

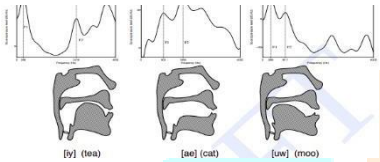
IDFT:

In this step, we are doing the inverse transform of the output from the previous step. Before knowing why we have to do

inverse transform we have to first understand how thesound produced by human beings.

The sound is actually produced by the glottis which is a valve that controls airflow in and out of the respiratory passages. The vibration of the air in the glottis produces the sound. The vibrations will occur in harmonics and the smallest frequency that is produced is called the fundamental frequency and all the remaining frequencies are multiples of the fundamental frequency. The vibrations that are produced will be passed into the vocal cavity. The vocal cavity selectively amplifies and damp frequencies based on the position of the tongue andother articulators. Each sound produced will have its unique position of the tongue and other articulators.

The following picture shows the transfer function of the vocal cavity for different phones.



The MFCC model takes the first 12 coefficients of the signal after applying the idft operations. Along with the 12coefficients, it will take the energy of the signal sample as the feature. It will help in identifying the phones. The formula for the energy of the sample is given below.

$$Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

Dynamic Features:

Along with these 13 features, the MFCC technique will consider the first order derivative and second order derivatives of the features which constitute another 26 features.

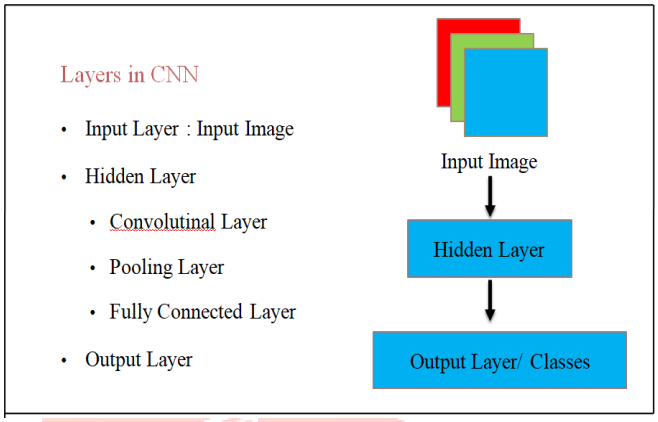
Derivatives are calculated by taking the difference of these coefficients between the samples of the audio signal and it will help in understanding how the transition is occurring.

So overall MFCC technique will generate 39 features from each audio signal sample which are used as input for the speech recognition model

E. Classification Using CNN

Convolutional neural network is the special type of feed forward artificial neural network in which the connectivity between the layers are inspired by the visual cortex. Convolutional Neural Network (CNN) is a class of deep neural networks which is applied for analyzing visual imagery. They have applications in image and video recognition, image classification, natural language processing etc. Convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel. Each input image will be passed through a series of convolution layers with filters (kernels) to produce output feature maps. Here is how exactly the CNN works.

Basically, the convolutional neural networks have 4 layers thatis the convolutional layers, ReLU layer, pooling layer and the fully connected layer.



Convolutional Layer

In convolution layer after the computer reads an image in the form of pixels, then with the help of convolution layers we take a small patch of the images. These images or patches are called the features or the filters. By sending these rough feature matches is roughly the same position in the two images, convolutional layer gets a lot better at seeing similarities than whole image matching scenes. These filters are compared to the new input images if it matches then the image is classified correctly. Here line up the features and the image and then multiply each image, pixel by the corresponding feature pixel, add the pixels up and divide the total number of pixels in the feature. We create a map and put the values of the filter at that corresponding place. Similarly, we will move the feature to every other position of the image and will see how the feature matches that area. Finally, we will get a matrix as an output.

ReLU Layer

ReLU layer is nothing but the rectified linear unit, in this layerwe remove every negative value from the filtered images and

replaces it with zero. This is done to avoid the values from summing up to zeroes. This is a transform function which activates a node only if the input value is above a certain number while the input is below zero the output will be zero then remove all the negative values from the matrix.

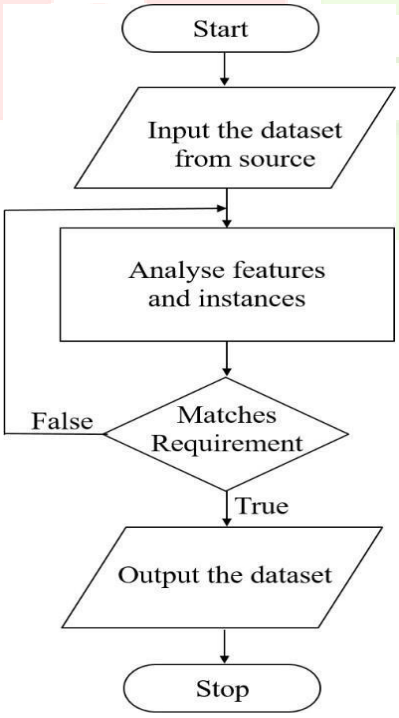
Pooling Layer

In this layer we reduce or shrink the size of the image. Here first we pick a window size, then mention the required stride, then walk your window across your filtered images. Then from each window take the maximum values. This will pool the layers and shrink the size of the image as well as the matrix. The reduced size matrix is given as the input to the fully connected layer.

Fully Connected Layer

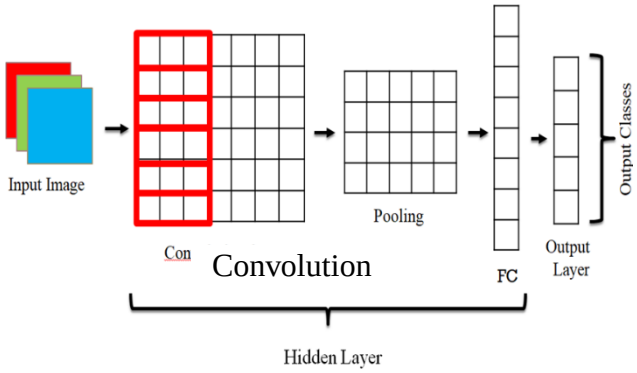
We need to stack up all the layers after passing it through the convolutional layer, ReLU layer and the pooling layer. The fully connected layer used for the classification of the input image. These layers need tobe repeated if needed unless you get a 2x2 matrix. Then at the end the fully connected layer is used where the actual classification happens.

The flowchart for Data Acquisition



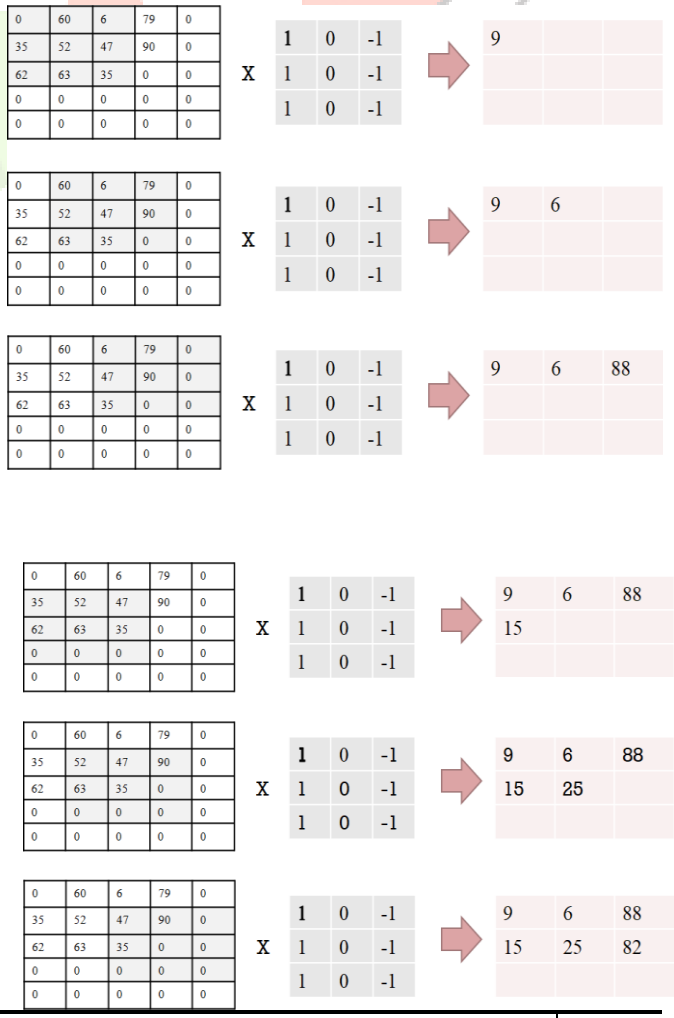
Typical CNN Architecture

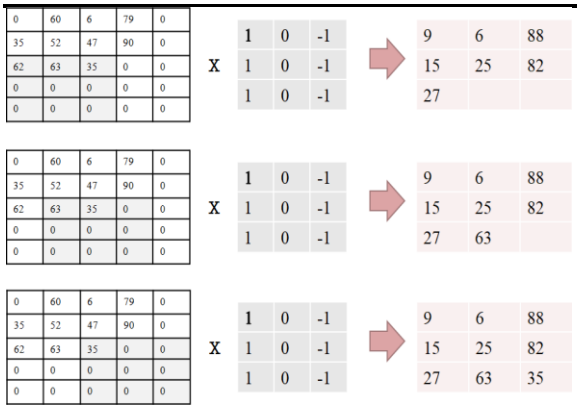
CNN architecture is inspired by the organization and functionality of the visual cortex and designed to mimic the connectivity pattern of neurons within the human brain. The

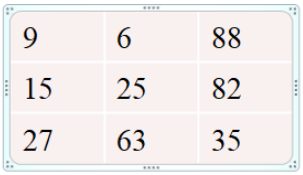


neurons within a CNN are split into a three-dimensional structure, with each set of neurons analyzing a small region or feature of the image. In other words, each group of neurons specializes in identifying one part of the image.

Convolutional Layer is the first step in CNN, here 3*3 partof the given matrix which was obtained from High-pass filter is given as input. That 3*3 matrix is multiplied withthe filter matrix for the corresponding position and their sumis written in the particular position. This is shown in the below figure. This output is given to pooling layer where thematrix is further reduced.



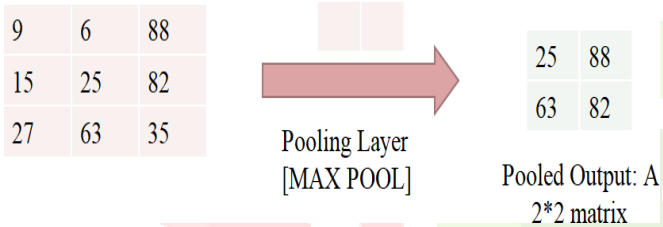




Convolutional Output: A 3*3 matrix

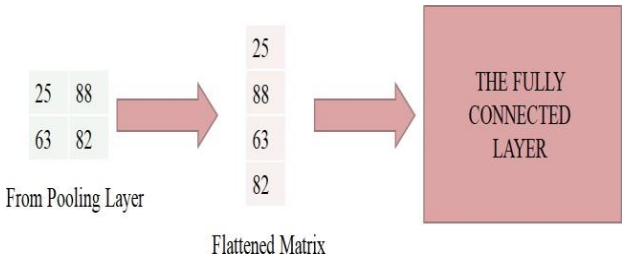
Convolution is followed by the rectification of negative values to 0s, before pooling. Here, it is not demonstrable, as all values are positive. In fact, multiple iterations of both are needed before pooling.

Pooling Layer

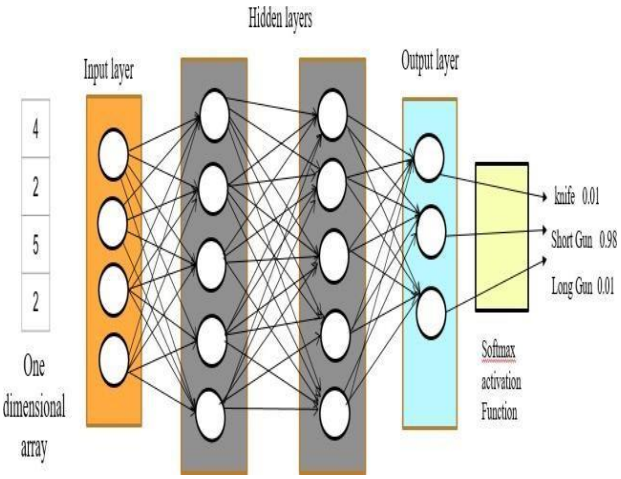


In Pooling layer 3*3 matrix is reduced to 2*2 matrix, this is done by selecting the maximum of the particular 2*2 matrix for the particular position. Figure 4.16 shows the Pooling Layer.

Fully connected layer and Output Layer



The output of the pooling layer is flattened and this flattened matrix is fed into the Fully Connected Layer. In the fully connected layer there are many layers, Input layer,

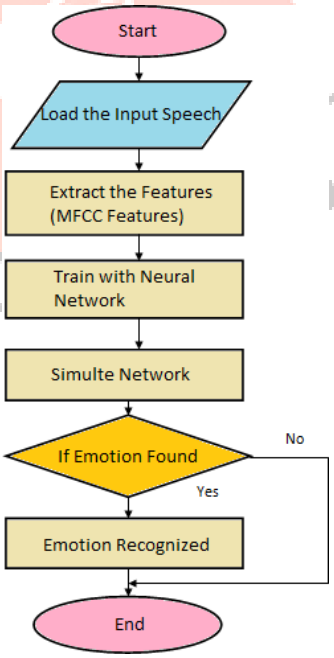


,

Hidden layer and Output layers are parts of it. Then thisoutput is fed into the classifier, in this case SoftMax Activation Function is used to classify the image into covid present or not.

Steps for Emotion Recognition

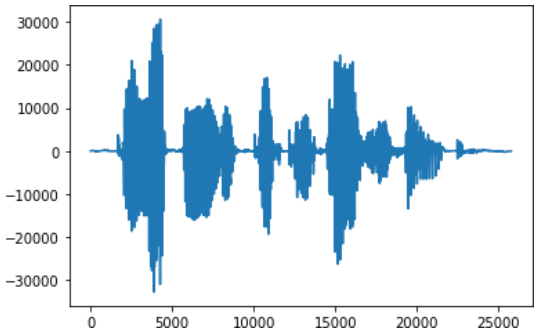
For research purpose, we have to divide our work intoseveral steps for the better processing and output of data. The flow diagram of those steps is shown in Figure below

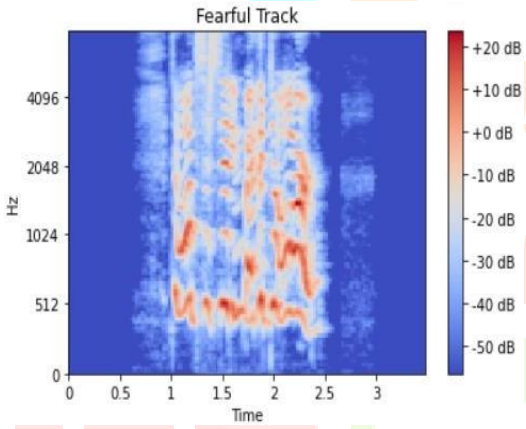


- In the first step, the input signal to be tested is loaded.
- The MFCC features of the voice sample are then released.
- Next step, input data is trained through the neural network.
- Then, network simulation is performed with CNN

IV EXPERIMENTS AND RESULT ANALYSIS

First, after testing the sample, the waveform of tested sample is generated. The waveform of one of the tested sample is shown in Figure.





V CONCLUSION AND DISCUSSION

The proposed method represents an outlook for recognizing emotion from human speech. This method has been enacted by the neural network. This dissertation mainly addresses the feature extraction which is proficient in the emotion recognition through speech. For the purpose of feature extraction, MFCC is used. The performance highly depends on the emotional speech samples. So, the speech samples from RAVDESS dataset should be taken properly for emotion recognition. The current accuracy for our proposed neural network model has the highest accuracy of 57.64%. In future,we will try to take care of the problems which clog the accuracy of the result, so that our project can be successful and we may have better accuracy.

The research over this idea fetches us the knowledge, that this technique is yet to play the vital role in medical and technical field. Speech emotion detection will play a wise role in upcoming equipment and systems. The authentication processes also highly lay their concern over this recognition formula. This idea may gain its assert over the vast field of computer science and other related branches by collecting lots of data on the predictable form and lay its root firm to become the indispensable one of the future world.

REFERENCES

[1] Md. Sham-E-Ansari, Shaminaj towfika Disha, atiqul Islam Chowdhury and Md. Khairul Hasan,"A neural network based approach for recognition of basic emotions from speech", 2020 IEEE Region 10 Symposium (TENSYMP), pp. 5-7, June 2020.

[2] R. Chen, Y. Zhou, and Y. Qian, "Emotion recognition using support vector machine and deep neural network," in Man-Machine Speech Communication (J. Tao, T. F. Zheng, C. Bao, D. Wang, and Y. Li, eds.), (Singapore), pp. 122–131, Springer Singapore, 2018.

[3] Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on hmm and svm," in 2005 International Conference on Machine Learning and Cybernetics, vol. 8, pp. 4898–4901 Vol. 8, Aug 2005.

[4] A. Badshah, J. Ahmad, N. Rahim, and S. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in 2017 International Conference on Platform Technology and Service (PlatCon), pp. 1–5, Feb 2017.

[5] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," pp. 333–336, 10 2017

[6] L. Fu, X. Mao, and L. Chen, "Relative speech emotion recognition based artificial neural network," in 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, vol. 2, pp. 140–144, Dec 2008

[7] Berlin Database of Emotional Speech http://emodb.bilderbar.info/docu/

[8] K. Rajvanshi, "An efficient approach for emotion detection from speech using neural networks", International Journal for Research in Applied Science and Engineering Technology, vol. 6, pp. 1062–1065, 05 2018.

[9] P. HarAar, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 137–140, Feb 2017.

[10] S. T. Saste and S. M. Jagdale, "Emotion recognition from speech using mfcc and dwt for security system," in 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 1, pp. 701–704, April 2017.

[11] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," IEEE Journal of Selected Topics in Signal Processing,vol. 11, pp. 1301–1309, Dec 2017.

[12] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–4, Dec 2016.

[13] X. Chen, W. Han, H. Ruan, J. Liu, H. Li, and D. Jiang, "Sequence- to- sequence modelling for categorical speech emotion recognition using recurrent neural network," in 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp. 1–6,May 2018.

[14] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," in ArXiv, Jan 2017.

[15] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. Mahjoub, "Speech emotion recognition: Methods and cases study," in 10th Inter- national Conference on Agents and Artificial Intelligence, pp. 175–182, 2018.

[16] F. Burkhardt, A. Paeschke, M. A. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in INTERSPEECH, 2005

[17] "Speech emotion recognition by adaboost algorithm and feature selection for speech emotion recognition by adaboost algorithm andfeature selection for support vector machines." http://www.academia.edu/2899315/Speech Emotion Recognition by AdaBoost Algorithm and Feature_Selection for Support Vector Machines

[18] [1] H. Hu, M. Xu, and W. Wu, "Gmm supervector based svm with spectral features for speech emotion recognition", 2007 IEEE International Conference on Acoustics, Speech and Signal ProcessingICASSP'07, vol.4, pp. IV–413–IV–416, April 2007.

[19] R. Chen, Y. Zhou, and Y. Qian, "Emotion recognition using support vector machine and deep neural network," in Man-Machine Speech Communication (J. Tao, T. F. Zheng, C. Bao, D. Wang, and Y. Li, eds.),(Singapore), pp. 122–131, Springer Singapore, 2018.

[20] Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on hmmand svm," in 2005 International Conference on Machine Learning and Cybernetics, vol. 8, pp. 4898–