CLASSIFYING VARIOUS EMOTIONS IN SPEECH USING LSTM

Mrs. K Sadhana[1], Siripuram Ramya[2], Yalla Lahari[3], Vemali Likitha[4], Chintalapati Shekar Varma[5]

[1]Associate Professor, [2,3,4,5]Students (B.Tech. CSE)
Lendi Institute of Engineering & Technology, AP, India

***Abstract*: Emotion recognition in speech is one of the main challenging components of Human Computer Interaction (HCI). Emotion detection has grown to be one of the maximum vital roles in customers that stumble on cutting-edge emotion of a person's emotion and suggest him the apt product or help him to improve the demand of the company or product. Where that is utilized in apps like Alexa, Cortana, Google Assistant and Siri. It has been changing the way of human beings interacting with the devices, homes, cars, jobs and also includes the huge database used to extract emotions, which contribute to speech emotion recognition and its limitation that relate to express such as happiness, anger, sad moods. We can use this technology in distinct fields consisting of security, medicine, entertainment, schooling.**

***Keywords*: Deep Learning, Human computer interaction, classification, speech data base, LSTM, Python, Librosa.**

1. INTRODUCTION

Today human is being interacting greater with the generation in place of peer-to-peer. This enables to enhance in generation which includes social media codecs which enables to speak virtually however now no longer emotionally. Due to this we cannot join or engage with the humans round us emotionally. As social abusing or crime rate has been increasing because of which humans are turning or cause do the unlawful sports which motive harm to the society. So, to identifying the emotions of people we have linked with the present scenario, which indicates what facilitates the interpretation of their emotions. Emotions can be expressed in a various way, along with facial features, physical signs, and language. Researchers has been using many techniques to discover the solution for detecting emotions in several of reading deep in artificial intelligence. So, we had been using the audio to extract feelings from information and making it use to set the rules which shows the feelings of speech.

SPEECH RECOGNITION

Speech recognition is a feature of devices that encounter phrases in a language which is spoken aloud in which we convert speech indicators into virtual indicators while working. To discover them we've many technologies which made simpler referred as 'Automatic Speech Recognition' utilized in Alexa, Cortana, Google assistant and Siri which are connected to our lives.

EMOTION RECOGNITION

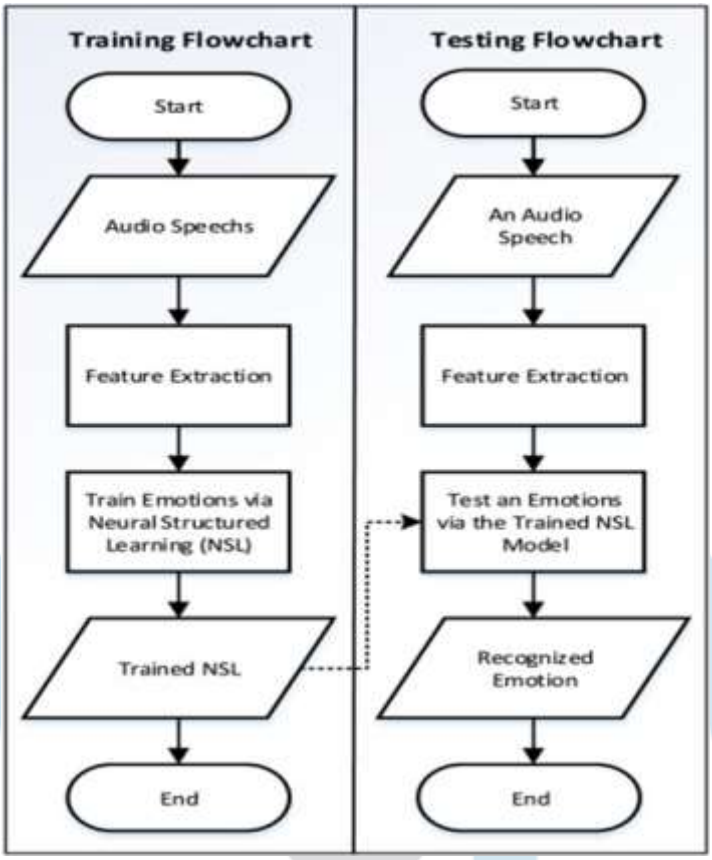
It is a system to discover the emotion of a human. Human has remarkable potential to discover feelings. But it's tough to discover the equal feelings for a gadget which we want to educate them and also tough to train machines in which it gets the information which incorporates the historical past noise. This is used in lots of fields including medical, education, fraud detection etc.

SPEECH EMOTION RECOGNITION

The location where speech emotion is identified is known as the speech recognition. Various surveys states that this kind of development performs a foremost function in growing the imminent technologies. It is a hard venture to decide the feelings primarily based totally on their nearby voice, which is ambiguous and shortage in availability of speech database. This feature can be used in many of the applications including security, medicine, entertainment, and schooling.

2. PROPOSED WORKS

Speech is the form of audio which should connect with the systems. There are many ways to become aware of the speech alerts to make prediction and become aware of them. In this project, we enforce LSTM.As audio documents which might be large in size. So, gadget must understand and become aware of the emotion accurately. The main goal is to make sure that it provide accurate results.Here we use the below flow chart to detect emotion of speech.



3. LITERATURE REVIEW

"Speech Emotion Recognition using Lstm", Nithya Roopa, S. Prabhakaran M, Betty.P have published in paper International Journal of Recent Technology and Engineering (IJRTE) has achieved 78.2% by using ravadess dataset. "Speech Emotion Recognition", R. Nihaal Datt Sai, Syed Shahbaaz, U. Bhanu Prakash which had been published in compilance engineering journal achieved an accuracy of 95% of using Ravadess dataset.

4. DATACOLLECTION

For every problem to solve we have to gather certain information in order to solve. We have a variety of techniques to collect the huge data sets which are available in internet resources. Majority of task is done due to available dataset. In this project, we have used Toronto emotional Speech Set (TESS) dataset. There are 2800 data points (audio files) in total, where each emotion consists of 200 files

OAF -Older actor study

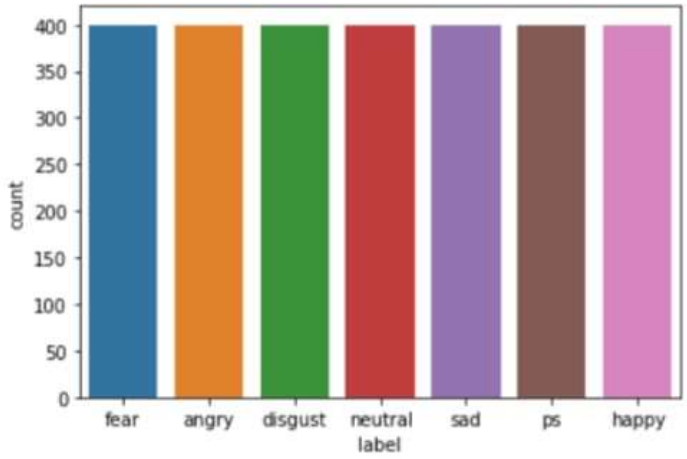
YAF – Younger actor study

SNO	EMOTION	NO OF FILES
1	OAF_Fear	200
2	OAF_Pleasant Suprise	200
3	OAF_ Sad	200
4	OAF_Angry	200
5	OAF_Disgust	200
6	OAF_Happy	200
7	OAF_Neutral	200
8	YAF_ Fear	200
9	YAF_ Pleasant Suprise	200
10	YAF_ Sad	200
11	YAF_ Angry	200
12	YAF_ Disgust	200
13	YAF_Happy	200
14	YAF_ Neutral	200
15	TOTAL	2800

Table 1. Audio dataset detail.

5. AUDIO PROCESSING

Audio files need to be pe-processed before extraction of features. This pre-process includes noise and silence removal. When we hear an audio clip with different background noises, it can be hard to understand what is being said. Human brains are good at analyzing sound frequencies and separating them into their individual components, helping us to focus on important sounds while ignoring background noise. The ability to split up a single frequency in an audio file can be achieved by a mathematical technique called the Fourier Transform. Speech is the form of audio which transforms a time-area signal into a frequency signals which ought to connect with the computer systems. We have learned that the amplitude and frequency of a signal change over time (independently of one another).

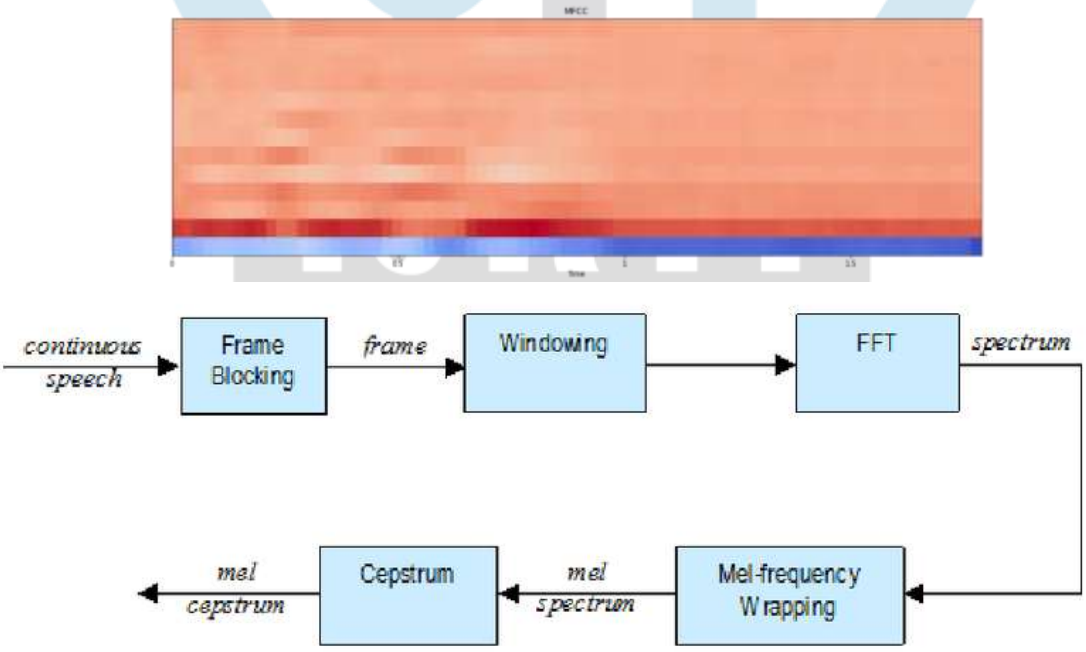


6. FEATURE EXTRACTION

Speech signals is depicted in terms of amplitude, frequency, pitch etc. It is difficulty to extract the feature from an individual speaker. There are many techniques to available for feature extraction, in this project we have using MFCC (Mel-frequency ceptral coefficients). It is not easy to extract to recognize audio from speech which are in waveforms because of huge variability.

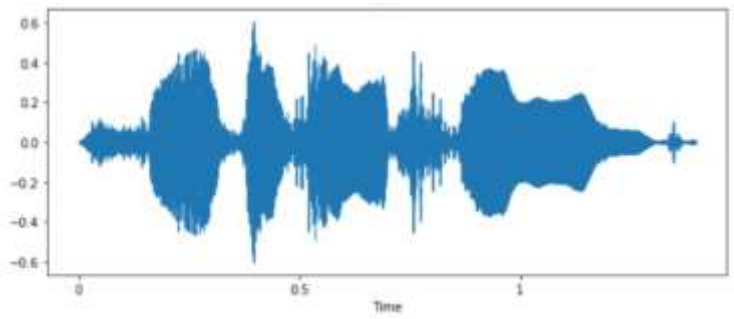
6.1 Mel-frequency ceptral coefficients (MFCC):

The signals of Mel Frequency Cepstrum Coefficient (MFCC) are a small feature set (usually around 10-20) that briefly describes the overall shape of the spectral begin.



Way to Extraxt features of MFCC

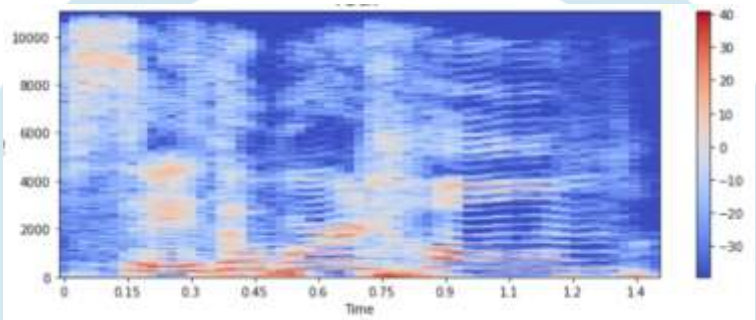
By printing the shape of mfccs we can get how many mfccs are calculated on how many frames.



Audio signals with resp. Time

6.2 MEL-SPECTOGRAM

This is a spectogram which is converted to a Mel-scale. The logarithmic shape of mel-spectogram allows us to recognize feelings higher due to the fact human beings understand sound in logarithmic scale. Therefore, the Mel-spectogram corresponds to the time vs. log-mel frequency illustration, which changed into received at some point of MFCC computation.



MEL- SPECTOGRAM

7. BUILDING THE MODEL

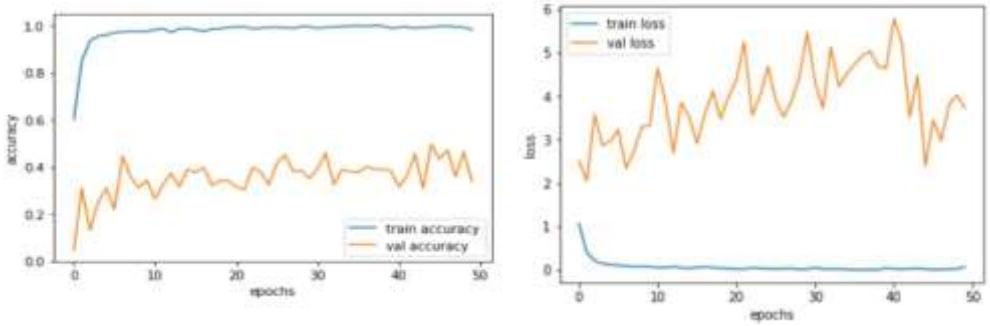
LSTM is sequential model (machine learning models that input or output sequences of data.) And we wrote this code in Python using the Keras library.

7.1 DEFINING NETWORK

LSTM is a framework which applied on several layers. Keras is neural network used to develop evaluating deep learning models which is series of layers which build stack of layers and arrange them in a sequence in which they should be interlinked and used to produce outcome.

7.2 EVALUATING NETWORK

Network need to be trained once. So, it helps to predict the performance of built model and metrics used in this model is accuracy of prediction.



From the first, graph we can say that trained accuracy is more when compare to validation accuracy. Similarly, in second graph we can say that trained loss has been decreasing compared to the validate loss. So, we can say that it is best model.

RESULT

An accuracy rate about 67% is achieved from the data model which classifies emotions using lstm model to predict emotions such as calm, angry, fear, happy, neutral, surprise, sad using TESS (Toronto Emotion Speech set) dataset.

REFERENCES

[1] Nithya Roopa S., Prabhakaran M, P Betty, "Speech Emotion Recogition using deep learning", published in International Journal of Recent Technology and Engineering (IJRTE) in 2018.

[2] R.Nihaal Datta Sai, Syed Shahbaaz, Bhanu Prakash, "Speech Emotion Recognition using lstm and rnn", published in compilance engineering journal.

[3] Bennilo Fernandes and Kasiprasad Mannepalli "Speech Emotion Recognition Using Deep Learning LSTM", Department of Electronics & Communication Engineering, Koneru Lakshmaiah Education Foundation, Guntur, AP, India.
[4] Speech emotion recognition using rnn, B. Siva Rama Krishna, N. Supriya, K. Nagamalleswara rao, G. Mounikad, S. Ajay Bharani "Speech emotion recognition using rnn" published in Turkish Journal of Physiotherapy and Rehabilitation; 32(3) ISSN 2651-4451 | e-ISSN 2651-446X.
[5] Speech emotion recognition methods by Babak Basharirad and Mohammadreza Moradhaseli. A literature review Cite as: AIP Conference Proceedings 1891, Published Online: 03 October 2017.
[6] Article Published in applied science, A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms Sung-Woo Byun and Seok-Pil Lee.