Speech Emotion Recognition*

1st Suhas Gogi
electronics and communication
KLE technological University

2nd Rahul H Hanumagatti
Electronics and communication
KLE technological University

3rd Asif Musafir
Electronics and Communication
KLE technological University

4th Marigouda Patil
Electronics and Communication
KLE technological University

***Abstract*—**
***Index Terms*—**

I. Introduction

Human speech contains many features that the listener learns to unpack the rich information given by the speaker. The speaker also in shares features such as tone,energy, speed, and other acoustic properties,which helps capture the subtext or intention and literal words.

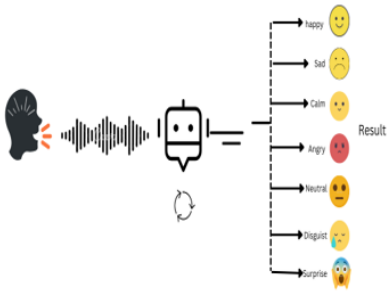

Fig. 1. image of speech emotion recognition.

The interaction between the humans and computers increases as the AI tries to understand the emotions that humans are going through .Speech is natural way for humans to express their emotions hence many emotion recognition applications work with speech. A very basic SER works by extracting features such as pitch frequency feature , special feature and energy related features. In particular, the speech emotion recognition task is one of the most important problems in the field of para linguistics. This field has recently broadened its applications, as it is a crucial factor in optimal human computer interactions, including dialog systems.

The goal of speech emotion recognition is to predict the emotional content of speech and to classify speech according to one of several labels (i.e., happy, sad, neutral, and angry).Various types of deep learning methods have been applied to increase the performance of emotion classifiers; however, this task is still considered to be challenging for several reasons. First, insufficient data for training complex neural network-based models are available, due to the costs associated with human involvement.

Identify applicable funding agency here. If none, delete this.

In this work, we propose a robust technique of emotion classification using speech features and transcriptions. The objective is to capture emotional characteristics using speech features, along with semantic information from text, and use a deep learning based emotion classifier to improve emotion detection accuracies. We present different deep network architectures to classify emotion using speech features and text. The main contributions of the current work are: - Proposed a CNN model for emotion classification using Speech features (MFCC, Spectrogram) - Proposed a CNN model for emotion classification using both speech features (MFCC, Spectrogram) and transcriptions

II. Ease of Use

A. PROPOSED METHODS

In this paper,we first extract the features from given data using the library given by python known as librosa .The features that we have extracted for this audio data path are mfcc,mel,chroma,contrast,spectrogram which together provide a deep neural network both semantic relationships and the necessary low-level features required to distinguish among different emotions accurately.Experiments have been performed on speech features independently to achieve accuracies greater than existing methods.

III. FEATURES WE EXTRACTED

Chroma :

- The chroma feature is a descriptor, which represents the tonal content of a musical audio signal in a condensed form.
- In audio file analysis, an audio file can consist of 12 different pitch classes.These pitch class profiles are very useful tools for analyzing audio files. The term

chromagram represents the pitches under an audio file, in one place so that we can understand the classification of the pitches in the audio files.

- Short Time Fourier Transforms and Constant Q Transforms are used for chroma feature extraction.
- One mainproperty of chroma features is that they capture harmonic and melodic characteristics of music, while being robust to changes in timbre

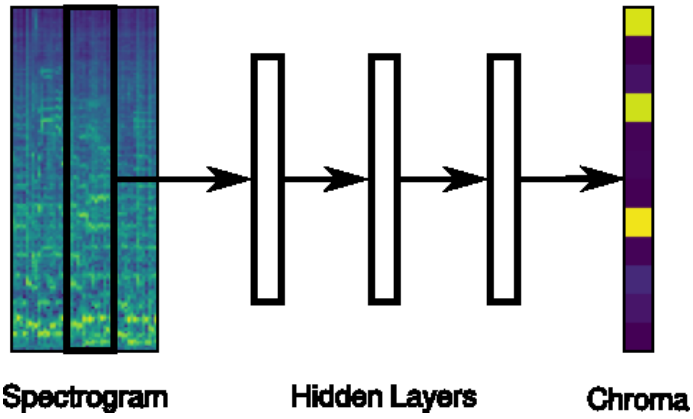

Fig. 2. Chroma feature.

Contrast :

- Most of the audio files have frequency whose energy changes with respect to that of time
- Spectral contrast is a way to anlayse energy of freaquency at each time stamp
- there is a difficulty in measuring the energies therefore we use contrast to measure that energy variations The above

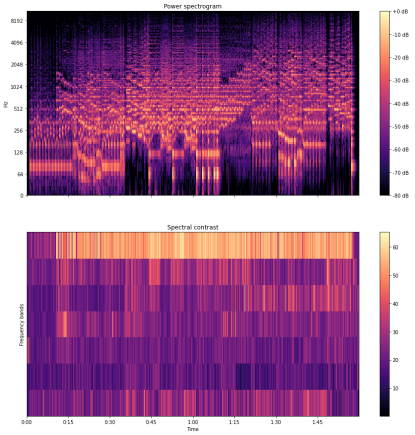

Fig. 3. contrast feature.

image represents spectral and power spectrum of audio in different time frames. The energy contrast is measured by comparing the mean energy in the peak energy frame to that of the bottom or valley energy frame.

Mel-Frequency Cepstral Coefficients (MFCCs):

- It Can be derived from a type of inverse Fourier transform representation.
- Short term power spectrum of any sound represented by the Mel frequency cepstral (MFC) and combination of MFCC makes the MFC.
- It can be derived by mapping the Fourier transformed signal onto the male scale using triangle or cosine overlapping windows.
- It can be derived by mapping the Fourier transformed signal onto the male scale using triangle or cosine overlapping windows.
- Where after taking the logs of the powers at each of the Mel frequencies and after discrete cosine transform of the Mel log powers give the amplitude of a spectrum. The amplitude list is MFCC.

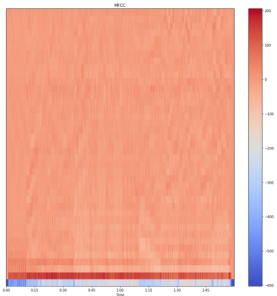

Fig. 4. MFCC feature.

IV. LSTM(LONG TERM SHORT MEMORY

An LSTM module has a cell state and three gates which provides them with the power to selectively learn, unlearn or retain information from each of the units. The cell state in LSTM helps the information to flow through the units without being altered by allowing only a few linear interactions. The LSTM network has three layers, an input layer, a single hidden layer followed by a standard feedforward output layer.

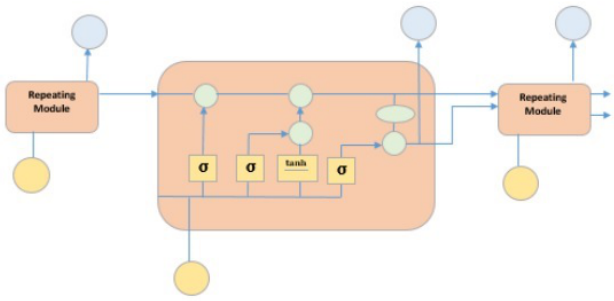

Fig. 5. image of speech emotion recognition.

The picture above depicts four neural network layers in yellow boxes, point wise operators in green circles, input in yellow circles and cell state in blue circles. An LSTM module has a cell state and three gates which provides them with the power to selectively learn, unlearn or retain information from each of the units. The cell state in LSTM helps the information to flows through the units without any changes

by allowing very minute amount of interactions. Every unit in this has a input,output and output gate which helps in adding or removing information to the cell state.The forget gate helps to forget the information from the previous cell which uses sigmoid function .The input gate controls the information flow to the current cell state using a point-wise multiplication operation of 'sigmoid' and 'tanh' respectively.Finally, the output gate decides which information should be passed on to the next hidden state

A. working of LSTM

LSTM network was designed to remove issue that we were facing with RNN(Recurrent Neural Network) due to vanishing gradient problem.They have feedback connections that make them different to more traditional feedforward neural networks.LSTM retains useful information about previous data in the sequence to help with the processing of new data points.As the result it is very good at processing speech ,text and genral time series

By giving an example we will try to make you understand LSTM model Consider we are trying to predict monthly ice cream sales. As one might expect, these vary highly depending on the month of the year, being lowest in December and highest in June.An LSTM network can learn this pattern that exists every 12 periods in time. It does not just use the previous prediction but rather retains a longer-term context which helps it overcome the long-term dependency problem faced by other models. It is worth noting that this is a very simplistic example, but when the pattern is separated by much longer periods of time (in long passages of text, for example), LSTMs become increasingly useful.
5mm

LSTM output at a particular point in time is dependant on three things:

- cell state - current long-term memory of network
- hidden state - output at the previous point in time
- The input data at the current time step

B. purpose of using LSTM model

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.

C. About the dataset

RAVDESS(The Ryerson Audio-Visual Database of Emotional Speech and Song): The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent.Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio only (16bit, 48kHz .wav), Audio Video (720p H.264, AAC 48kHz, .mp4), and Video only (no sound). Note, there are no song files for Actor 18.Speech file (Audio Speech Actors01-24.zip, 215 MB) contains 1440 files: 60 trials per actor x 24 actors = 1440.Video files are provided as separate zip downloads for each actor (01-24, 500 MB each), and are split into separate speech and song downloads:

Each of the 7356 RAVDESS files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.mp4). These identifiers define the stimulus characteristics: Modality (01 = full-AV, 02 = video-only, 03 = audio-only). Vocal channel (01 = speech, 02 = song). Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised). Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion. Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"). Repetition (01 = 1st repetition, 02 = 2nd repetition).Actor (01 to 24. Odd numbered actors are male, even numbered actors are female). In this data set we are only going to use the audio data that is present in the file.

D. Evaluation and Discussion

E. Emotion classification on standard data set

in agreement with previous research efforts we give an effective methods for emotion recognition with our benchmark results on ravdess data set. We have split the training and testing data by normal means .We have displayed the emotion classification results in table1 as shown

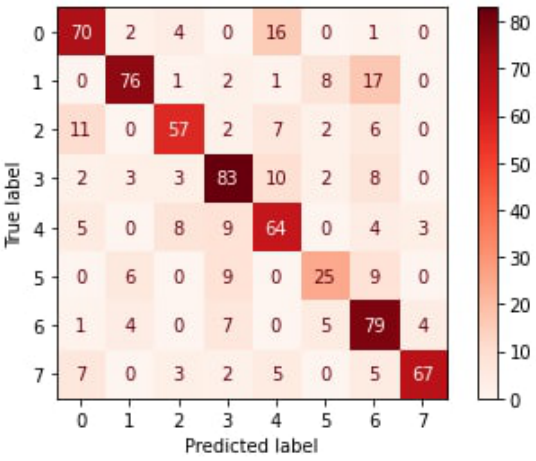

Fig. 6. Confusion matrix.

(1)

	precision	recall	fi-score	support
0	0.73	0.75	0.74	93
1	0.84	0.72	0.78	105
2	0.75	0.67	0.71	85
3	0.73	0.75	0.74	111
4	0.62	0.69	0.65	93
5	0.6	0.51	0.55	49
6	0.61	0.79	0.69	100
7	0.91	0.75	0.82	89
accuracy			0.72	725
macro avg	0.72	0.7	0.71	725
weighted avg	0.73	0.72	0.72	725

Fig. 7. Confusion Matrix table.

F. Methodology diagram

we have formed the fallowing block diagram to understand the the method that we have used in a easy way As seen in

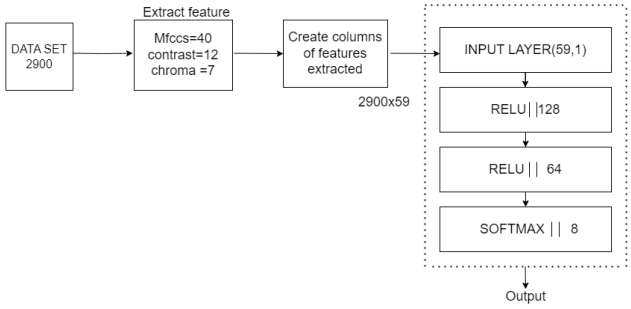

Fig. 8. Methodology diagram.

the block diagram we have taken a total of 2900 audio data files as the input.We then extracted the features from audio files file the features that we have extracted are mfcc,chroma and contrast .We then created proper columns to store the Extracted features

there is a single input layer of 59 layers, 2 RELU layers of 128 layers and then it is made to 8 layers by softmax.When classifying upon a sequence usually we stack some LSTM returning sequences, then one LSTM returning a point, then Dense with softmax activation. Is it possible instead to give the last non-sequential LSTM a softmax activation