Speech Emotion Recognition using Deep Learning Techniques

Omkar Patil (CS00104), Pravin Lokhande (CS00084), Pratik Chougule (CS00102)

Guide: Prof. P. D. Lanjewar

ABSTRACT

Human Speech is one of the important factors for Human-Computer Interaction (HCI). Many techniques have been utilized to extract emotions via speech emotion recognition (SER). Previously, traditional techniques were implemented in SER, but recently Deep learning techniques are being used as an alternative. This paper overviews some Deep Learning Algorithms, find out their outputs and the drawbacks in those techniques.

I. INTRODUCTION

As human beings speech is amongst the most natural way to express ourselves. We depend so much on it that we recognize its importance when opting to other communication forms like emails and text messages where we often use emojis to express the emotions associated with the messages. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today’s digital world of remote communication.

Emotion detection is a challenging task, because emotions are subjective. There is no common procedure on how to measure or categorize them. We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions present in them. This system can be useful in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis. In this study, we attempt to detect underlying emotions in recorded speech by analyzing the features of the audio data of recordings.

TABLE 2. Summarized form of some acoustic variations observed based on emotions.

Emotions	**Pitch**	**Intensity**	**Speaking rate**	**Voice quality**
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

II. PROBLEM STATEMENT

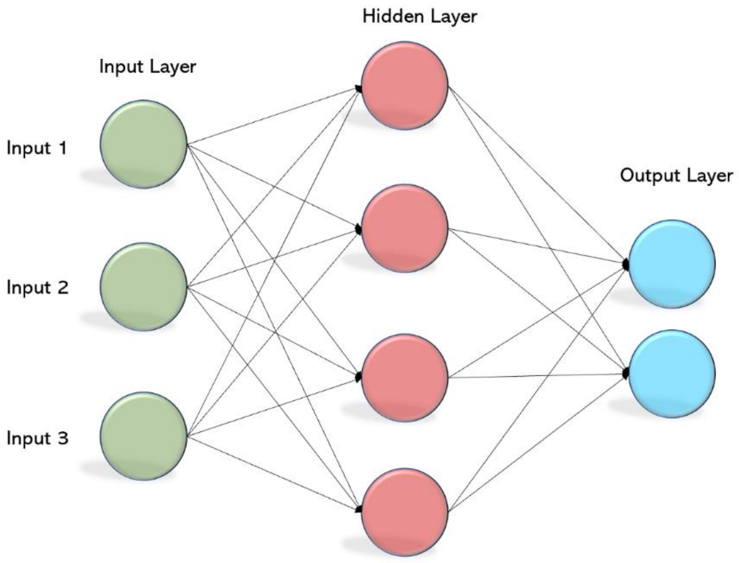
To create a platform which helps to predict the emotion of the users from their speech.

III. METHODOLOGY

We have implemented three Deep Learning Algorithms namely MLP (Multi-Layer Perceptron), CNN (1D-Convolutional Neural Network), and LSTM (Long Short-term Memory i.e., modified Recurrent Neural Network) with specific parameters. In the proposed system, we have used various features of audio data such as MFCC (Mel-Frequency Cepstral Coefficient), Chroma, Mel Spectrogram, Zero Crossing Rate, Root Mean Square Value. We are using RAVDESS and TESS datasets for training our model.

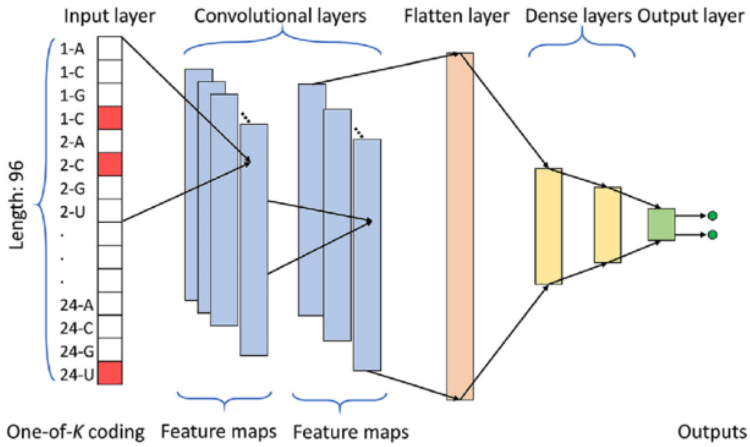
a. MLP Classifier:

A multilayer perceptron (MLP) is a fully connected class of feedforward artificial neural network (ANN). Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result.



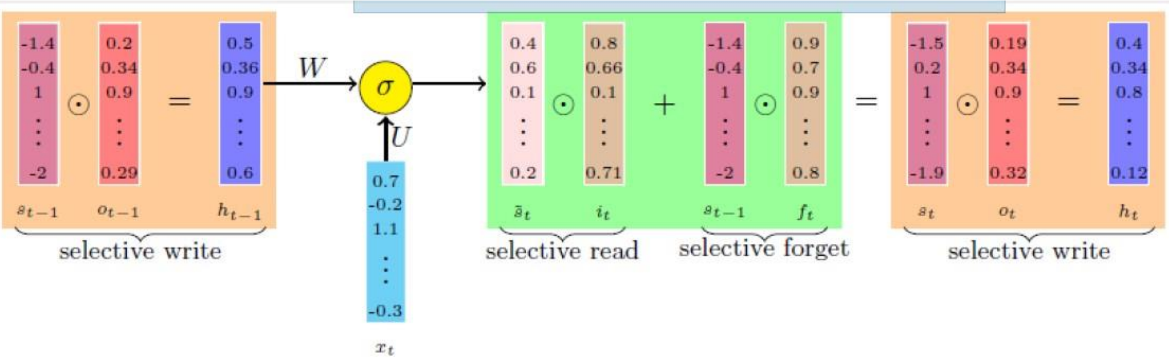
b. CNN:

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. Here we are using 1D CNN where kernel slides linearly along the 1D data/ array and learns the features. It is similar to 2D CNN and consists of similar layers as 2D CNN like convolutional layer, pooling layer, etc. We use fully connected layer to obtain required output.



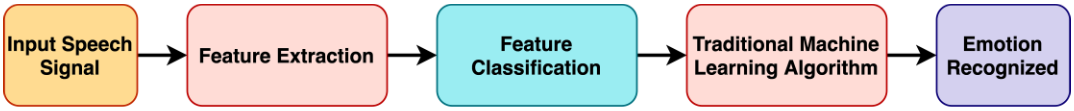
c. LSTM:

It is modified RNN. It uses gates to control flow of information. It uses three methods namely selective write, selective read and selective forget.

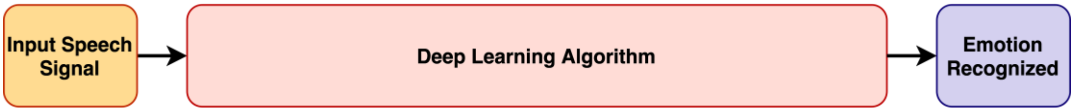


It decides how much previous information is important using selective write and uses a vector o(t-1) to decide the fraction to be passed. Hidden layer is created using previous state and o(t-1). It uses hidden state to create current temporary state (i.e., st^). Now using selective read the important information is only extracted and further use for building new current state. But the whole information is not directly used instead an input gate is used to control the information flow. Now using selective forget least useful data of previous state is excluded, but the whole in not vanished instead a fraction of each data is passes using forget gate. Thus, selective read and selective forget together form the next input.

IV. FLOWCHART



Traditional Machine Learning Flow Mechanism



Deep Learning Flow Mechanism

V. CONCLUSION

Various investigations and surveys about Emotion Recognition, Deep learning techniques used for recognizing the emotions are performed. It is necessary in future to have a system like this with much more reliable, which has endless possibilities in all fields. Deep learning can be used effectively to predict the emotions of the users. This model will be beneficiary in certain sectors and will be user-friendly. Using the proposed model, we can classify number of emotions such as anger, neutral, sad, calm, happy, disgust, fear and surprise.

VI. REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Communications of the ACM, vol. 61, no. 5, pp. 90–99, 2018.

[2] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," Information Fusion, vol. 49, pp. 69–78, 2019.

[3] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," IEEE Access, vol. 5, pp. 326–337, 2017.