International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

SPEECH EMOTION RECOGNITION

Darshan K.A[1], Dr. B.N. Veerappa[2]

[1]U.B.D.T. College of Engineering, Davanagere, Karnataka, India
[2]Dr. B.N. Veerappa, Department of Studies in Computer Science and Engineering, U.B.D.T. College of Engineering, Davanagere.

Abstract – *The aim of this paper is to document the development of speech emotion recognition system using CNN. A model is designed that could recognize the emotion in a speech sample. Various parameters are modified to improve the accuracy of the model. The paper also aims to find out the factors which are affecting the accuracy of the model and key factors which are needed to improve the efficiency of the model. The paper concludes with a discussion on accuracy of different CNN architecture and parameters needed to improve accuracy and possible areas of improvement.*

Key Words: **ASR, Deep Neural Network, Convolution Neural Network, MFCC.**

1. INTRODUCTION

Speech is a means of communication in humans. Speech is series sequence of a words of pre-established language and it is an essential medium for communication. Speech technology is a computing technology that empowers an electronic device to recognize, analyse and understand spoken words or audio. There are many researches and studies going on from several years in understanding the building blocks of human brain that makes up human intelligence. Human brain is a complex organ which has inspired for exploration in Artificial Intelligence. The aim of Artificial intelligence is to develop a system that are intelligent and capable enough to develop the behaviour and thoughts like human behaviour and thoughts. One of the important fields of research in AI and machine learning is Automatic speech recognition (ASR), which aims to design machines that can understand the human speech and interact with people through speech.

In order to communicate effectively with people, the systems need to understand the emotions in speech. Therefore, there is a need to develop machines that can recognize the paralinguistic information like emotion to have effective clear communication like humans. One important data in paralinguistic information is Emotion, which is carried along with speech. A lot of machine learning algorithms have been developed and tested in order to classify these emotions carried by speech. The aim to develop machines to interpret paralinguistic data, like emotion, helps in human-machine interaction and it helps to make the interaction clearer and natural. In this study Convolution Neural Networks are used to predict the emotions in speech sample. To train the network RAVDEES and SAVEE dataset are used which contains speech samples that were generated by the 24 actors and 4 actors respectively. The results showed accuracy of **77%.**

2. SYSTEM DESIGN

Convolution Neural Networks (CNN's) are used to differentiate the speech samples based on their emotion. Databases such as RAVDEES and SAVEE are utilized to prepare and assess CNN models. Kears (TensorFlow's high-level API for building and training deep learning models) is used as the programming framework to implement CNN models. Seven exploratory arrangement of the present work are explained in this section.

2.1 Database

▪ **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)**
The Ryerson Audio-Visual Database of Emotional Speech and Songs contains 24 professional actors (12 females, 12 male), articulating two lexically-similar sentence in a neutral North American accent. Speech samples include emotions such as calm, happy, sad, angry, fearful, surprise, and disgust. Each emotion is produced at two stages of emotional intensity (normal, strong), with an additional neutral emotion.

▪ **Survey Audio-Visual Expressed Emotion (SAVEE)**
The audio folder of this dataset consists of speech samples recorded by four male speakers. For each emotion class there are 15 sentences. The starting letter of the file name represents emotion class. The letters 'su', 'sa', 'n', 'h', 'f', 'd' and 'a' represent 'surprise' 'sadness' 'neutral', 'happiness', 'fear', 'disgust', and 'anger' emotion classes respectively. For example, 'a03' represents the third anger sentence.

2.2 Pre-processing

The first step involves organizing the audio files. The emotion in an audio sample can be determined by the unique identifier of the file name at the 3rd position, which represents the type of emotion. The dataset consists of five different emotions

1. Calm 2. Happy 3. Sad 4. Angry 5. Fearful

2.3 Defining Labels

Based on the number of classes to classify the speech labels are defined. Some of the classes are as follows:

Class: positive and negative
Positive: Calm, Happy.
Negative: Fearful, Sad, Angry.
Class: Angry, Sad, Happy, Fearful, Calm.
Class: Angry, Sad, Happy, Fearful, Calm, Neutral, Disgust, surprised.

2.4 Feature Extraction

The Shape of the Speech signal determines what sound comes out. If the shape is determined accurately, then the correct representation of the sound being generated is obtained. The job of Mel Frequency Cepstral Coefficients' (MFCC's)is to correctly represent it. MFCCs is used as input feature. Loading and converting audio data into MFCCs format is done by python package librosa.

2.5 Designing the Dimensions of the Model

This step includes designing the layers of the CNN model, choosing the activation function, choosing the appropriate pooling options, defining the SoftMax to classify the speech into different classes.

2.6 Model Training and Testing

The model is trained with the training dataset and tested with the test data set. Actual values are compared with the predicted values. This comparison gives us the accuracy of the model.

2.7 Architecture of CNN

In the current study, the deep neural network architecture actualized is convolutional neural network. In the proposed architecture after each convolutional layer max-pooling layer is placed. To establish non linearity in the model, for activation function Rectified Linear Units (ReLU) is used in both convolutional and fully connected layers.

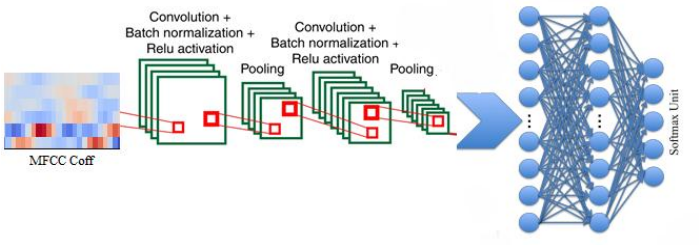

Fig 1: Architecture of CNN

Batch normalization is used to improve the firmness of neural network, which normalizes the result of the preceding activation layer by reducing the number by what the hidden unit values move around and allows each of the layer in a network to learn by itself. Dense layer is used; in which all the neurons in a layer are connected to neurons in the next layer and it is a fully connected layer. SoftMax unit is used to compute probability distribution of the classes. The number of SoftMax to be used depends on number of classes to classify the emotions. The model took between 10hrs to 14hrs to be trained. CPU's consumes lot of time to train the model, instead of that GPU's can be used to speed the training process. The several cores in GPU accelerate the speed and saves much time. The Figure 1 shows the Convolutional Neural Network architecture used in this paper. Lighter CNN architecture is also used to classify among a greater number of classes and good results are obtained.

3. RESULTS AND DISCUSIONS

The experiment is carried out with deeper CNN architecture. Batch normalization, max pooling, ReLU activation function are used. The model is trained for 250 epochs. For training 70% of dataset is used and for testing 30% of dataset is used. The model is trained to classify among the two different classes of male voice i.e., male happy and male sad. For this model the accuracy was 87% which is a very good result. The below figures show the emotion distribution and confusion matrix of it.

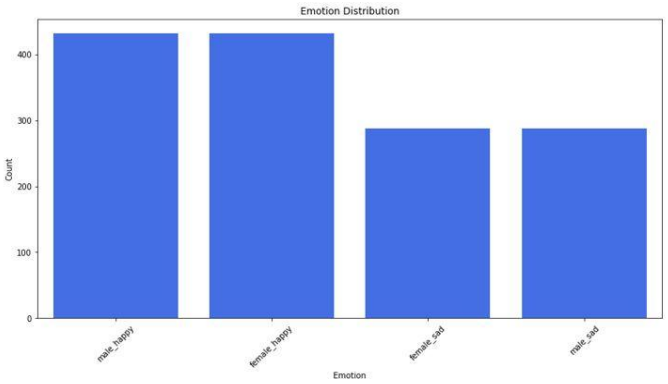

Fig 2:Distribution of Emotion in Dataset

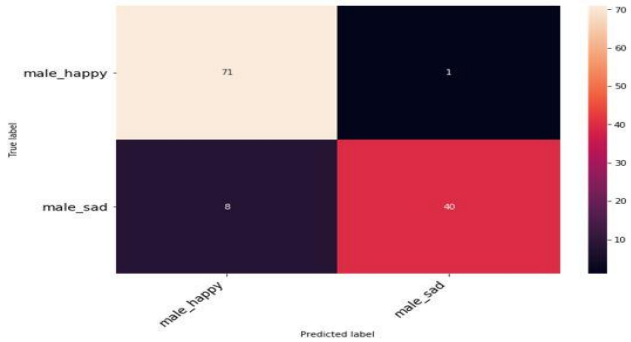

Fig 3: Confusion Matrix for First Experiment

The second experiment was carried out with lighter CNN architecture. ReLU activation function and max pooling are used. Two datasets are used to train the model (RAVDEES and SAVEE dataset). The model is trained for 1000 epochs. For training 80% of dataset is used and for testing 20% of dataset is used. The model is trained to classify among the ten different classes of male a and female voice. For this model

the accuracy is **77 %** which is a very good result. The below figures show the confusion matrix of it.

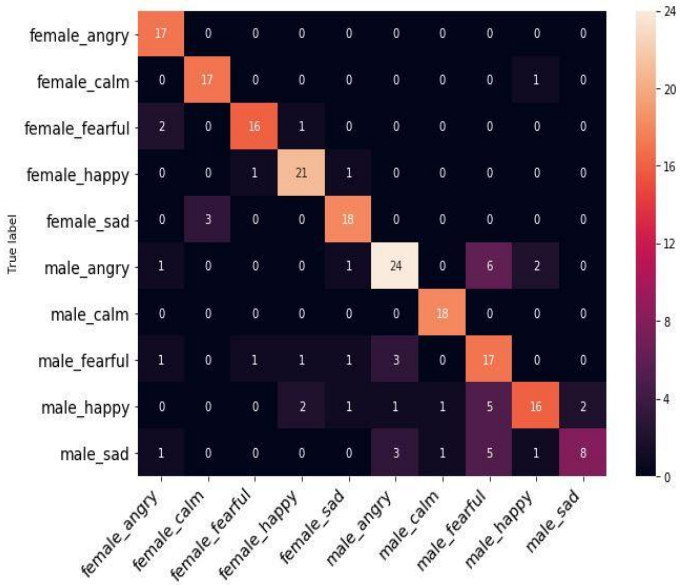

Fig 4: Confusion Matrix for Second Experiment

3.1 Real Time Analysis

The user selects the speech sample to find out the emotion in it. The model extracts the MFCC features from the sample and predicts the emotion in the class as per the pre-defined emotion class.

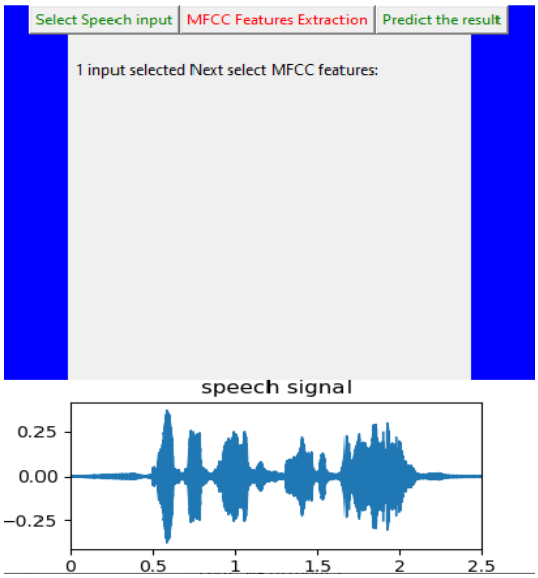

Fig 5: User Selects a Speech Sample

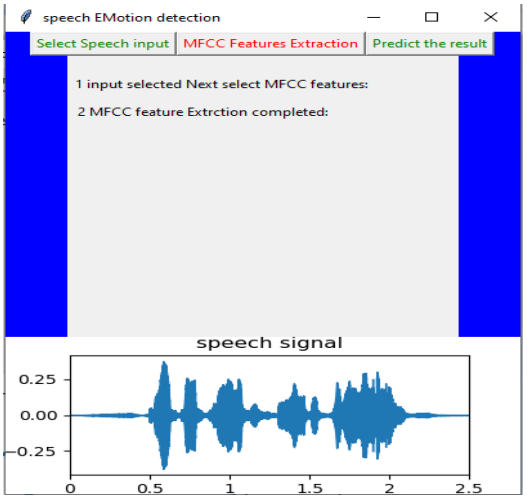

Fig 6: MFCC features Extraction from a Speech Sample

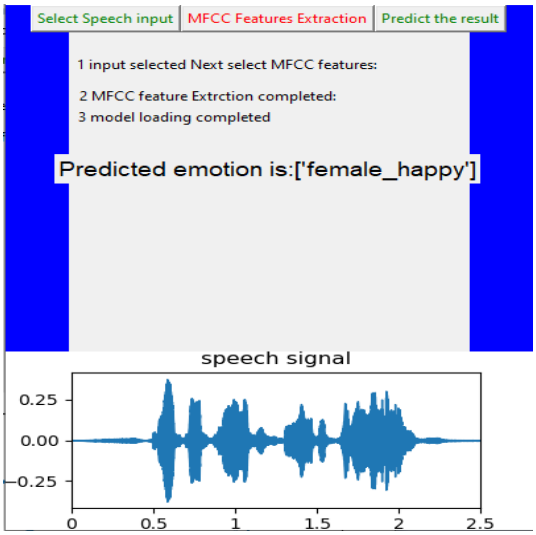

Fig 7: Predicted Emotion for a Speech Sample

4. CONCLUSION AND FEATURE WORK

Various experiments are conducted by changing the several parameters like dimension of the model, number of epochs, changing the partition ratio between training and test data set. Different accuracy was found for different experiments. A lighter CNN architecture with 80% of training data and 20% of test data gave good result compared to deeper CNN architecture while classifying among ten classes. The accuracy of this model was found to be 77%. The performance of deeper CNN model was found to be very good when the classification was among two classes the reason is there were a greater number of training samples available to classify among two classes. When the same model was used to classify among ten classes the training dataset was divided into ten labels this led to the fewer number of training samples available for each class. This led to the poor accuracy of the model. With the greater number of training samples available for each class and with the help of GPU's to speed up the training process more accuracy can be achieved in the future enhancements.

REFERENCES

[1] Arianna Mencattini, Eugenio Martinelli, Giovanni Costantini, Massimiliano Todisco, Barbara Basile, Marco Bozzali, and Corrado Di Natale. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowledge-Based Systems, 63:68–81, 2014.

[2] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, pages 312–317. IEEE, 2013.

[3] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 16(8):2203–2213, 2014.

[4] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Towards real-time speech emotion recognition using deep neural networks. In Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on, pages 1–5. IEEE, 2015.

[5] WQ Zheng, JS Yu, and YX Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, pages 827–831. IEEE, 2015.

[6] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Bj¨orn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pages 5200–5204. IEEE, 2016.

[7] Andrew Ng. Sequence models. Deeplearning.ai on Coursera, February 2018.

[8] Michalis Papakostas, Evaggelos Spyrou, Theodoros Giannakopoulos, Giorgos Siantikos, Dimitrios Sgouropoulos, Phivos Mylonas, and Fillia Makedon. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. Computation, 5(2):26, 2017.

[9] Tom M Mitchell et al. Machine learning. wcb, 1997.

[10] C Bishop. Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn. Springer, New York, 2007.

[11] RS Suton and Andrew G Barto. Reinforcement learning: an introduction. adaptive computation and machine learning, 2002.

[12] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning, volume 112. Springer, 2013.

[13] Aur´elien G´eron. Hands on machine learning with scikit-learn and tensorflow, 2017.