International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Emotion Recognition of Manipuri Speech using Convolution Neural Network.

Gurumayum Robert Michael, Aditya Bihar Kandali

Abstract: *over the recent years much advancement are made in terms of artificial intelligence, machine learning, human-machine interaction etc. Voice interaction with the machine or giving command to it to perform a specific task is increasingly popular. Many consumer electronics are integrated with SIRI, Alexa, cortana, Google assist etc. But machines have limitation that they cannot interact with a person like a human conversational partner. It cannot recognize Human Emotion and react to them. Emotion Recognition from speech is a cutting edge research topic in the Human machines Interaction field. There is a demand to design a more rugged man-machine communication system, as machines are indispensable to our lives. Many researchers are working currently on speech emotion recognition(SER) to improve the man machines interaction. To achieve this goal, a computer should be able to recognize emotional states and react to them in the same way as we humans do. The effectiveness of the speech emotion recognition(SER) system depends on quality of extracted features and the type of classifiers used . In this paper we tried to identify four basic emotions: anger, sadness, neutral, happiness from speech. Here we used audio file of short Manipuri speech taken from movies as training and testing dataset . This paper use CNN to identify four different emotions using MFCC (Mel Frequency Cepstral Coefficient)as features extraction technique from speech.*

Keywords **: CNN,** ***emotion recognition, Human Machine interface, MFCC,.***

I. INTRODUCTION

Speech is the statement of one's sentiments or contemplations by verbalized sounds. Speech signal contains the data about speaker, language, message and feelings . Emotion is the expression of human feelings. It may be conveyed through face, movement or speech . Emotions are vital for passing on significant information.. speech contains different kind of emotions like happiness ,sadness ,fear, disgust, anger surprise etc. A detailed survey on speech emotion recognition (SER) is given in [1] which discussed about the features, classifier schemes and databases. Emotion detection of Assamese speech using Gaussian Mixture Model (GMM) classifiers and Mel frequency Cepstral co-efficient (MFCC) are described in[2]. A method of emotion classification for speech using short time long frequency power co-efficient (LFPC) and a discrete hidden Markov model (HMM) as the classifier is described in [3]. The proposed system yields 78% of accuracy and discussed classification of 6 emotions.[5]discuss about the emotion recognition from speech using MFCC and DWT for security system using SVM classifier. In [6] the authors have effectively utilized for speech-based emotion identification. They planned and trained the network for the detection of 6 essential emotions from speech. In this paper we tried to identify for basic emotion using CNN form Manipuri Speech.

Manuscript received on April 02, 2020.
Revised Manuscript received on April 20, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
G.R.Michael*, Dept. of ECE, Dibrugarh University Dibrugarh , India. Email: roberteld008@gmail.com
Dr Aditya Bihar Kandali., Electrical Department, Jorhat Engineering college, Jorhat, India. Email: abkandali@rediffmail.com

II. . MEL FREQUENCY CEPSTRAL COEFICIENT

Mel Frequency Cepstral Coefficient (MFCC) is the most generally utilized feature extraction method utilized in automatic speech recognition. MFCC relies upon human hearing acknowledgments which can't perceive frequencies over 1Khz.. Fig. 1 shows the complete steps to get the MFCC coefficient.

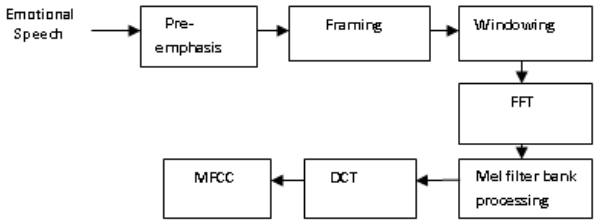

Fig.1 MFCC block

A. Pre-emphasis

In the speech spectrum more energy is concentrated at lower frequencies. Pre-emphasis increase the signal energy. Speech signal is passed through a filter which increases the signal energy.

$$Y[n] = X[n] * 0.95\ X\ [n-1] \qquad [1]$$

B. Framing and windowing

In this process speech signal is segmented into 20ms (Frame blocking) then, windowing is done by Hamming using the formula:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi}{N-1}\right), 0 \le n \le N-1 \\ 0, otherwise \end{cases} \qquad [2]$$

W(n) = Hamming window
N=number of input Samples
n= sample input index in time domain

C. FFT

Fourier transform converts each N samples of each frame from the time domain into the frequency domain. FFT is used to find all frequencies present in the particular frame [7].

D. Mel filter bank

The mel filter bank consists of 20- 30 triangular filters applied to each frame . The mel scale filter bank identify how much energy exists in a particular frame[7] .The given equation converts the normal frequency to the Mel scale :

F(Mel) = [2595* log 10 [1+F] 700] [3]

E. Discrete cosine transforms

DCT transforms the cepstral of the frequency domain into a coefficient namely quefrency domain. The result of this process is MFCC .

Fig 2. Shows the MFCC plot of different emotions.

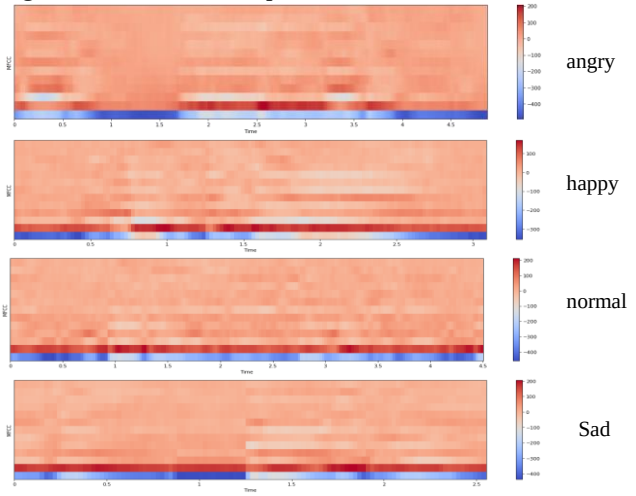

Fig 2. MFCC plot of different emotions

III. DATABASE

In this paper we have selected four emotion—anger, happiness, sadness, neutral. The steps for construction of our emotional speech database is shown in Fig. 3.

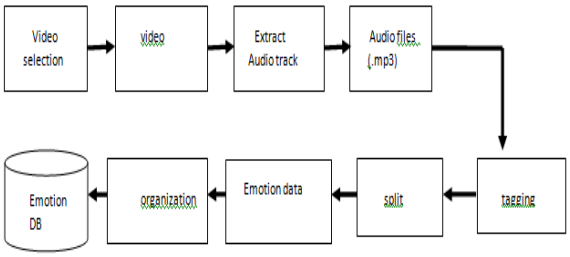

Fig.3 Emotional speech database

First step is to choose plays ,movies or clip that contain the emotional dictionary. The background sound is also incorporated with the video information for each of the selected items. The Audio is extracted for the selected videos using VLC player and the extracted audio information are saved as .mp3 audio files. The Tagging step is done by using Audacity, a free open-source digital audio processing software. Each tagged emotional area determine information like the start and end time, gender of the speaker, emotion, the noise, the title, and the speech all of which are passed together as a script file (including wav audio data) to the Split step. The Split step, separate the long audio files into emotional data units according to labeling data, and short audio files are altered manually. Finally, the isolated emotional data units are Organized as audio files with all the labeled data encoded in the file names e.g. **act2_a01.wav** indicates the speaker is actor no.2 ,male, the **_a** indicates angry and **01** indicates sl.no of angry database. **Fact1_h01.wav** indicate **F** for female actor ,h for happy and 01 is the sl. No.

Using this technique, we have collected 500 samples of emotional data from five actor, 3 male and 2 female. Out of 500 emotional speech database 400 samples are used consisting of 100 anger,100 neutral , 100sadness, 100 happiness samples for this experiment.

IV. CONVOLUTION NEURAL NETWORK :

We have used a CNN architecture derived from an image processing CNN**.** Figure 4 contains the block scheme of the architecture.

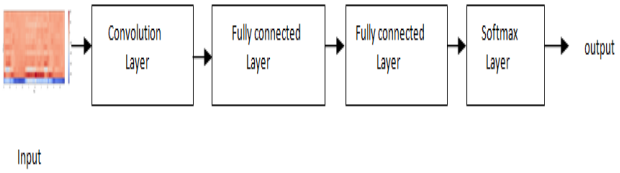

Fig.4. CNN architecture

The particularity of the CNN is the nearness of sets of convolutional and pooling layers. A convolutional layer has the reason to remove the organized data with submatrices filters(strides) parsing on the two-dimensional information. A pooling layer summarizes the output of the convolution matrix by aggregating the values of the stride submatrix into a single value. The CNN architecture includes also a number of dense (fully connected) layers, with the final (top) layer that contains the classifier [6]

TABLE 1. SUMMARY OF THE MODEL

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 216, 256)	2304
activation_1 (Activation)	(None, 216, 256)	0
conv1d_2 (Conv1D)	(None, 216, 256)	524544
batch_normalization_1 (Batch	(None, 216, 256)	1024
activation_2 (Activation)	(None, 216, 256)	0
dropout_1 (Dropout)	(None, 216, 256)	0
max_pooling1d_1 (MaxPooling1	(None, 27, 256)	0
conv1d_3 (Conv1D)	(None, 27, 128)	262272
activation_3 (Activation)	(None, 27, 128)	0
conv1d_4 (Conv1D)	(None, 27, 128)	131200
activation_4 (Activation)	(None, 27, 128)	0
conv1d_5 (Conv1D)	(None, 27, 128)	131200
activation_5 (Activation)	(None, 27, 128)	0
conv1d_6 (Conv1D)	(None, 27, 128)	131200
batch_normalization_2 (Batch	(None, 27, 128)	512
activation_6 (Activation)	(None, 27, 128)	0
dropout_2 (Dropout)	(None, 27, 128)	0
max_pooling1d_2 (MaxPooling1	(None, 3, 128)	0
conv1d_7 (Conv1D)	(None, 3, 64)	65600
activation_7 (Activation)	(None, 3, 64)	0
conv1d_8 (Conv1D)	(None, 3, 64)	32832
activation_8 (Activation)	(None, 3, 64)	0
flatten_1 (Flatten)	(None, 192)	0
dense_1 (Dense)	(None, 4)	772
activation_9 (Activation)	(None, 4)	0

Total params: 1,283,460
Trainable params: 1,282,692
Non-trainable params: 768

The model was implemented using Keras model-level library with the TensorFlow backend . programming was done in Python using karas .
The design chose for our application incorporates 8 convolutional channel, with activation ReLu , a maximum pooling and has 216 x 265 networks as input. The last layer comprises of a straightening and a dense layer of 192 neurons, followed by the classifier (Table 1 shows the summary of the model).

V. RESULT AND DISCUSSION

For this experiment we used a system setup consist of Core i3 intel core 2.0 GHz Processor, ASUS of 1 TB memory space, NVIDIA GeForce Nvidia GeForce 930M Graphics Card with windows 10 installed. For deep learning we used Tensor Flow 2.0 for implementing CNN model.
The proposed CNN model was implemented using TensorFlow. The training process was run for 1000 epochs with a batch size set to Initial learning rate was set to 0.00001 with a decay of 1e-6. Fig.5 shows the loss model graph. An accuracy of 46 % was accomplished per spectrogram. It is essential to see here that the general precision is low. This might be a direct result of the constrained dataset accessible for preparing the model. Fig.5 and Fig. 6 shows loss function and the confusion matrix predicted emotions vs true emotions respectively.

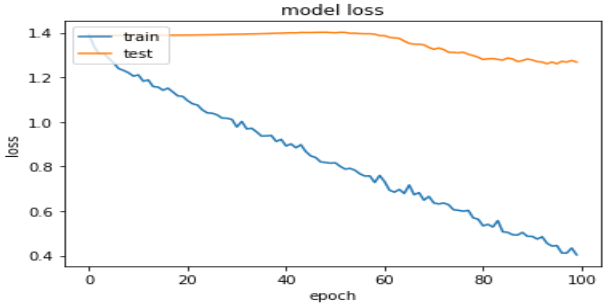

Fig.5 loss function

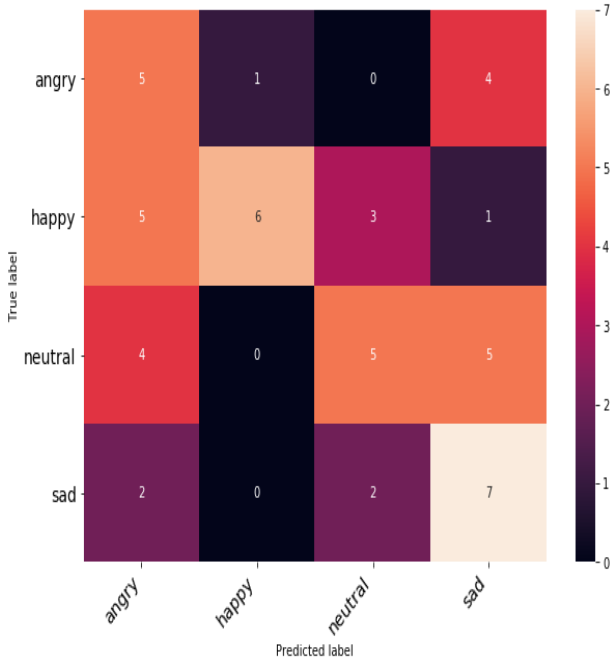

Fig. 6 Confusion Matrix

TABLE 2. f-SCORE OF DIFFERENT EMOTIONS

	precision	recall	f1-score	support
angry	0.31	0.50	0.38	10
happy	0.86	0.40	0.55	15
neutral	0.50	0.36	0.42	14
sad	0.41	0.64	0.50	11
accuracy			0.46	50
macro avg	0.52	0.47	0.46	50
weighted avg	0.55	0.46	0.47	50

VI. CONCLUSION

Different examinations and overviews about Emotion Recognition, Deep learning methods utilized for emotion recognition are performed. This experiment used CNN for solving emotion recognition problem, and a new database is constructed using Manipuri films and dramas for carrying out this experiment. Model is trained using TensorFlow 2.0. Accuracy rate of about 46% is achieved. TABLE 2. Shows the accuracy and f-score of different emotions. The next aim is to increase our database by collecting more real time data and improve the accuracy to 70% to 80%.

REFERENCES:

1. Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. “Survey on speech emotion recognition: Features, classification schemes and databases.” Pattern Recognition, 44(2011):572–587, ScienceDirect
2. Aditya Bihar Kandali, Aurobinda R and Tapan K. Basu,” Emotion recognition from Assamese speeches using MFCC features and GMM classifier” TENCON 2008 - 2008 IEEE Region 10 Conference, pp:1 – 5,2008
3. T. L. New, S. W. Foo and L. C. De Silva, “Speech emotion recognition using hidden Markov models”, Speech Commun., vol. 41, pp. 603–623, 2003
4. Laishram Rahul; Salam Nandakishor; L. Joyprakash Singh; S. K. Dutta, Design of Manipuri Keywords Spotting System using HMM , 2013 Fourth National Conference on Computer Vision,PatternRecognition, (NCVPRIPG)DOI:10.1109/NCVPRIPG.2013.6776249
5. Sonali T. Saste, Prof. S. M. Jagdale, Emotion Recognition from Speech Using MFCC and DWT for Security System, International Conference on Electronics, Communication and Aerospace Technology ICECA 2017
6. E. Franti, I. Ispas, V. Dragomir, M. Dascalu, Z. Elteto; I. Stoica, “Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots” in Romanian Journal of Information Science and Technology, vol. 20, Issue 3, pp. 222-240, 2017
7. Saikat Basu, Jaybrata Chakraborty, Arnab Bag, Md. Aftabuddin “A Review on Emotion Recognition using Speech” International Conference on Inventive Communication and Computational Technologies (ICICCT 2017)

AUTHORS PROFILE

Gurumayum Robert Michael ,Assistant professor in Dept. of ECE,Dibrugarh University Dibrugarh, Assam India. M.tech in Electronics design and Technology,Tezpur University. Research Area :Emotion Detection from speech, Speech recognition,Iot

Dr Aditya Bihar Kandali ,Associate Professor, Electrical Department, Jorhat Engineering college ,Assam,India. Phd,IIT kharagpur. Research Area: speech Recognition, Emotion Detection from speech,