iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XI **Month of publication:** November 2022

DOI: https://doi.org/10.22214/ijraset.2022.47463

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

Emotion Recognition from Manipuri Language Using MFCC and Convolution Neural Network.

Gurumayum Robert Michael[1], Dr. Aditya Bihar Kandali[2]
[1]*Assistant Professor, Department Of ECE, Dibrugarh University, Dibrugarh, Assam 786004, India*
[2]*Professor, Department of Electrical, Jorhat Engineering College, Jorhat, Assam 785007, India*

Abstract: ***Emotion recognition is a significant part in designing a better Human- Computer Interaction (HCI) System. Increasing applications of Emotion recognition from speech has been found in Education, gaming, medicine and automobiles industries. And after the Covid-19 pandemic identification of human emotions and action upon is the need of the hour. In this paper we try to identify four emotions by training a CNN model using Manipuri speech data set. We test our model using real time speech signal. Our model identifies sad emotion accurately but fails to identify angry emotion.***
Keywords: emotion recognition, CNN, HCI, MFCC, SER.

I. INTRODUCTION

Emotion is one of the significant components that express individuals' perspective. Human voice gives data about whether the speaker is happy or sad without looking at the person. The emotion state ingrained in the voice is a significant component that can be used alongside language and text for effective communication.

Two kinds of mediums have been recognized for human to human communication [1]: Explicit- that convey information through substantial message like language and Implicit- that conveys certain information, like emotions, that cannot be conveyed through explicit message. For more robust human-computer interaction (HCI), the computer should be able to perceive emotional state similarly as we human do. For more human like interaction with machine an HCI system should have both explicit message and implicit emotional delivery.

[2] Earlier we train a CNN model using Manipuri Speech data set. An accuracy of 46% was achieved in the model and an accuracy of 71 % is achieved using data augmentation. A 25% improvement is obtained by using combination of augmented and synthetic data. The accuracy measurement that we recorded above is based on the same dataset. In this report we applied a new recorded audio in our model and tried to predict the emotions. We input 10 audio signals in our model and predictions have been done. We observed how our model performs once we apply it to a completely new dataset with different audio quality, speaker and background noises. Our model does well in predicting the "sad" emotion but it has difficulty in predicting "angry "emotion.

II. MFCC

MFCC is a prevailing feature extraction method used in automatic speech recognition(SER).The coefficients are more sturdy and give great outcomes for noisy conditions.

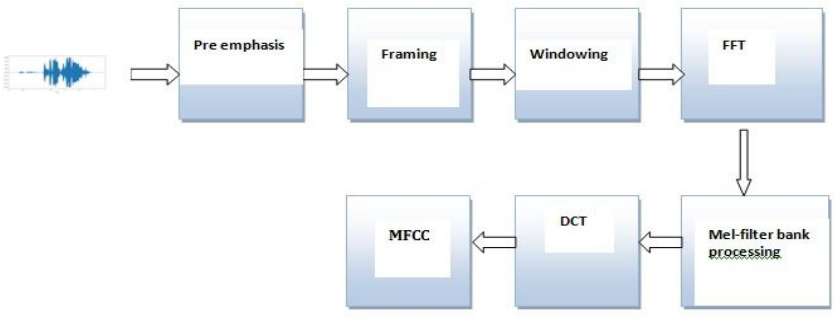


Fig.1 MFCC blocks

A. Pre-emphasis

To amplify the high frequencies on the signal Pre-emphasis is used. Since higher frequencies have smaller magnitude it balance out the frequency spectrum . It enhances the signal-to-noise ratio (SNR). [3]The most widely used pre emphasis is first order system given by the equation (1):

$$y[n] = x[n] * 0.95\ x\ [n-1] \qquad (1)$$

B. Framing and Windowing

Signals processing is done for a speech segment called frames of sizes typically between 20 and 40 ms [4].
The frames are overlapped to smooth out the transitions by a predetermined size. To minimize the signal discontinuities and spectral distortion windowing is done in each frame Hamming window is used for this purpose, which is given by:

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

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

C. FFT

Fast fourier transform(FFT) converts the frequency spectrum of each frame. Each N sample is converted from time domain in to frequency domain . All the frequencies in a frame can be determined by using FFT. To implement discrete fourier transform over a signal x(n), FFT is used which is given by[3] :

$$X_k \sum_{n=0}^{N-1} x(N) e^{\frac{-j2\pi kn}{N}}, k = 0,1,2,3.....N-1 \qquad (3)$$

D. Mel Filter bank

Mel filter bank is a set of 20 to 30 overlapped triangular filters. It determine the energy that exist in a frame.
Mel scale is linearly spaced below 1kHz Hz and spaced logarithemically above 1kHz

$$m(f)=2595*\log10(1+ f/700) \qquad (4)$$

E. Discrete Cosine Transforms

The cepstral in the frequency domain is transformed into a coefficient, specifically quefrency domain. The outcome of this operation is MFCC

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

III. CNN

CNN is gaining momentum in the area of deep learning in specially in audio domain. It is a dominant model used in text and image recognition. CNN is also gaining attention in computer vision applications. The basic architecture of CNN is shown in Fig.2

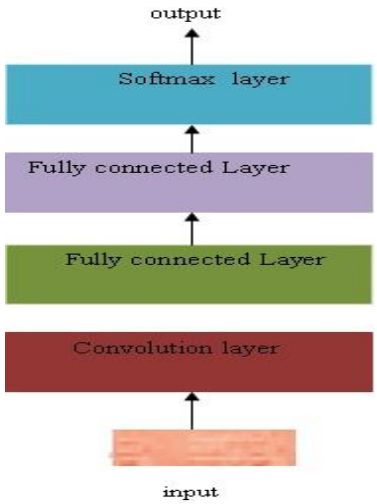


Fig.2 CNN architecture

There are two essential parts of convolution neural network: a) Feature extraction b) classifier .After the feature extraction from the audio signal it is passed to a classifier to mapped the signals into their proper classes.
Our model is implemented using Keras model-level library with TensorFlow backend. Our model consist of 8 convolution channel with ReLu activation. It has 216x265 networks as input and the last layer is a dense layer or 192 neurons , followed by the classifier.

IV. RESULT AND DISCUSSION

We ran the prediction for all the ten recorded audio file and the results is recorded in the given table no.1. We observed that our model identify sad emotion accurately but fails to identify angry emotion. The variation in the result may be due various reasons like the background noise or inaccurate emotion acted by our speaker while our model is trained on data from professional actor. Fig.3 shows MFCC plot of different emotions and Some of the predicted emotions are shown in Fig.4 .Further improvement has to be done and more real time data has to be collected to test our Model. The real challenges faced in collecting the data are to record real time emotional data.

Sl .No	Audio No.	Input/Actual emotion	Predicted emotion
1	Audio 1	Angry	Neutral
2	Audio 2	sad	sad
3	Audio 3	happy	Neutral
4	Audio 4	sad	sad
5	Audio 5	angry	neutral
6	Audio 6	happy	happy
7	Audio 7	sad	sad
8	Audio 8	neutral	neutral
9	Audio 9	angry	happy
10	Audio 10	happy	happy

Table 1. ACTUAL/INPUT EMOTION VS PREDICTED EMOTIONS FOR EACH AUDIO INPUT.

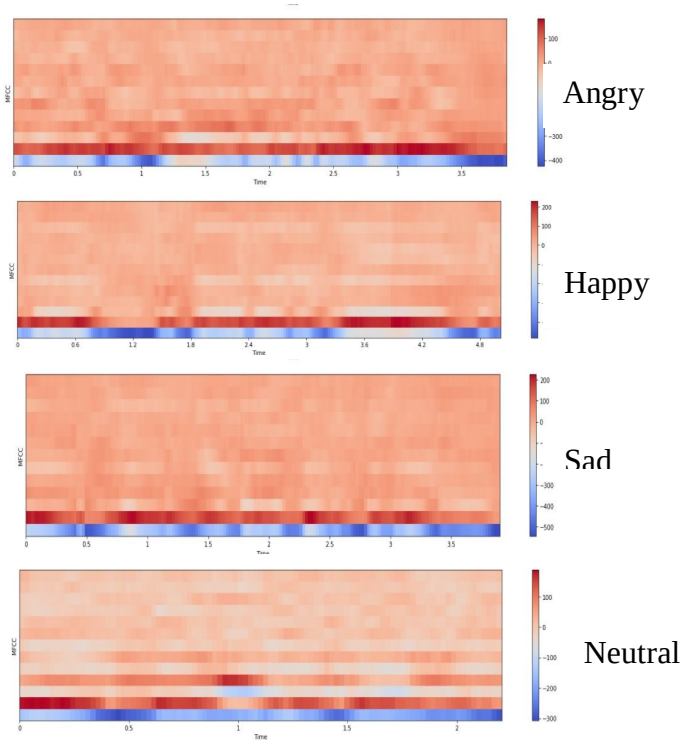


Fig.3 MFCC plots

```
filename = '/kaggle/input/labels/labels'
infile = open(filename,'rb')
lb = pickle.load(infile)
infile.close()

# Get the final predicted label
final = newpred.argmax(axis=1)
final = final.astype(int).flatten()
final = (lb.inverse_transform((final)))
print(final) #emo(final) #gender(final)
```

```
['sad']
```

```
filename = '/kaggle/input/labels/labels'
infile = open(filename,'rb')
lb = pickle.load(infile)
infile.close()

# Get the final predicted label
final = newpred.argmax(axis=1)
final = final.astype(int).flatten()
final = (lb.inverse_transform((final)))
print(final) #emo(final) #gender(final)
```

```
['neutral']
```

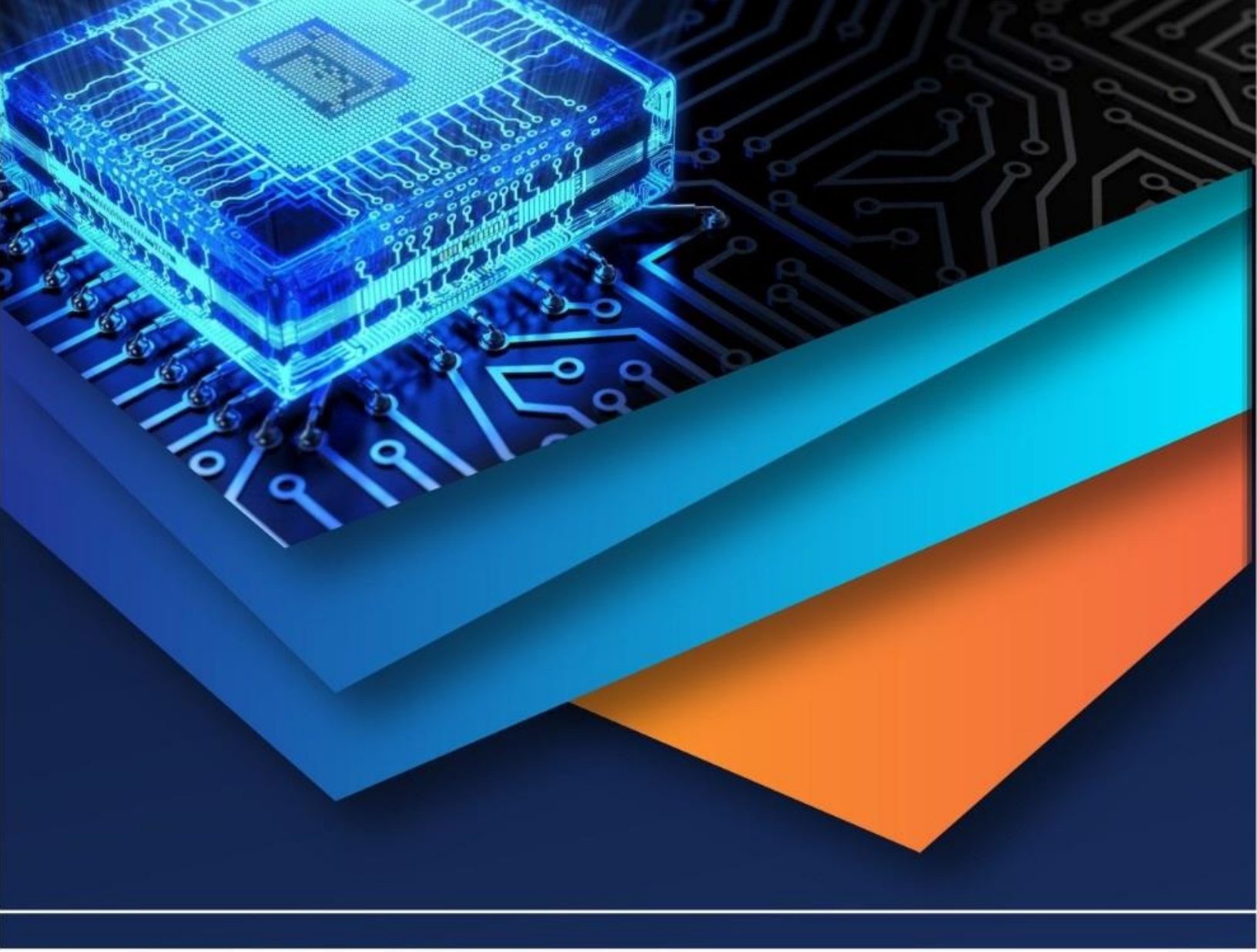
Fig.4 Predicted Emotions

V. CONCLUSION

In this paper we trained a CNN model using the Manipuri Speech data set. An accuracy of 46% was achieved in the model and an accuracy of 71 % is achieved using data augmentation. A 25% improvement is obtained by using a combination of augmented and synthetic data. We also observed that our model identifies sad emotion accurately but fails to identify angry emotion..Further improvement has to be done and more real time data has to be collected to test our Model. The real challenges faced in collecting the data are to record real time emotional data. The future work is to be train the model using more real time data to improve the accuracy.

REFERENCES

[1] References I.Hong, Y.Ko,H.Shin,Y.Kim,"Emotion Recognition from Korean Language using MFCC, HMM, and Speech Speed",ISSN 1975-4736,MITA2016
[2] G.R.Michael,A.B.Kandali,"Emotion recognition of Manipuri speech using convolution Neural Network.",International journal od recent technology and Engineering,Vol-9,Issue 1,May 2020
[3] https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html
[4] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," Journal of Computing, vol. 2, no. 3, pp. 138–143, Mar. 2010
[5] M.S. Likitha,1 Sri Raksha R. Gupta,2 K. Hasitha3 and A. Upendra Raju4," "Speech Based Human Emotion Recognition Using MFCC ",International conference on wireless communication, signal processing and networking(WiSPNET), pp.2257-2260, March 2017
[6] Keras: The Python Deep Learning library, : https://keras.io



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY