AUDIO AUGMENTATION FOR EMOTION RECOGNITION USING MANIPURI SPEECH

Gurumayum Robert Michael[1], Dr. Aditya Bihar Kandali[2]

[1]Department of Electronics and communication, Dibrugarh University, Dibrugarh, Assam (India)

[2]Department of Electrical Engineering, Jorhat Engineering College, Jorhat, Assam (India)

ABSTRACT

The abstract should summarize the content of the paper. Try to keep the abstract below 200 words. Do not make references nor display equations in the abstract. Data augmentation is an approach to improve the quality of training data, void over fitting and improve the ruggedness of the models. In this paper we want to compare the performance of an emotion recognition system using data augmentation technique and compare the loss function with and without the data augmentation. We work on four basic emotions states such as happy, neutral, sad and angry from Manipuri speech. The audio file used is extracted from short Manipuri speech taken from YouTube videos and dramas and used for training and testing dataset. We use CNN model to identify various emotions. Features extraction from speech is done using MFCC (Mel Frequency Cepstral Coefficient). An accuracy of 46% was achieved in the model (without data augmentation) 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.

Keywords: Emotion recognition, CNN, Data Augmentation, MFCC, Manipuri speech

I. INTRODUCTION

Data augmentation is a technique which tries to increase the data amount and acoustic variety. The aim of data augmentation is to increase performance on robustness of the ASR system [3][4][5].One of the major problems faced by Speech recognition researcher is the lack of data. In this paper our objective is to compare alternative solution to lack of data.Our result demonstrates that some data augmentation or speech synthesis technique works well to improve emotion speech recognition for low resource language.In this paper a 25% improvement is obtain by using combination of augmented and synthetic data. To generate syntactic data for audio, we apply noise injection, changing pitch and speed.

II. DATA AUGMENTATION FOR AUDIO

To generate syntactic data for audio, we can apply noise injection, shifting time, changing pitch and speed. Fig1. Shows the original wave plot of a speaker and different data augmentation are performed on the same sound and Fig

2. Shows Spectrograms plots of some of the above augmentation:

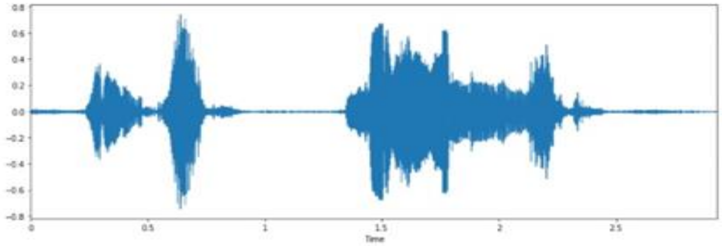


Fig.1 (a). Original sound

A. Noise Injection:

This audio augmentation method do is to add static noise in the background. Fig. 1. (b). Shows the wave with the added noise.

$$y(n) = \alpha + x(n) \qquad (1)$$

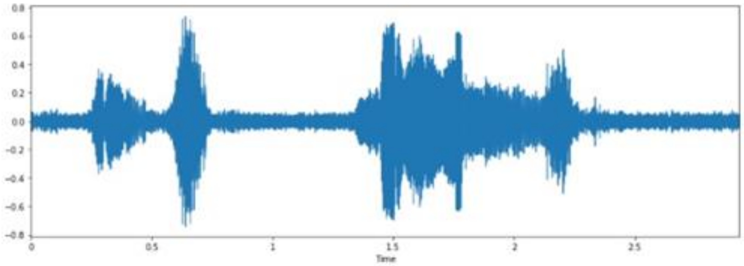


Fig.1.(b). With added noise

B. Stretch

This one is one of the more dramatic augmentation methods. The technique in a real sense extends the Audio. So the span is longer, yet the sound wave gets stretched as well. Consequently presenting and impact that sounds like a slow movement sound

$$y(n) = \begin{cases} x\left(\frac{n}{M}\right) & M = 0 \pm M, \pm 2M \ldots\ldots. \\ 0 & otherwise \end{cases} \qquad (2)$$

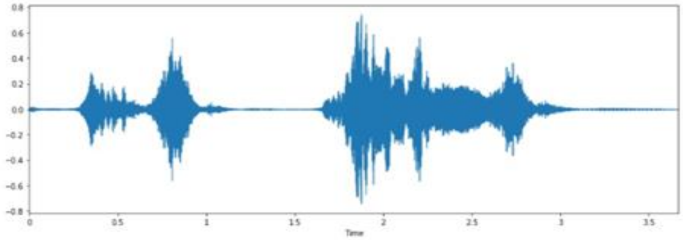


Fig.1(c).Stretch

C. Time shifting

Time Shifting moves the audio randomly to either the left or right direction, within the fix audio duration. Comparing with the original plot, the audio wave pattern is exactly same, other than a tiny bit of delay before the speaker starts speaking.

y(n)=x(n-N) (3)

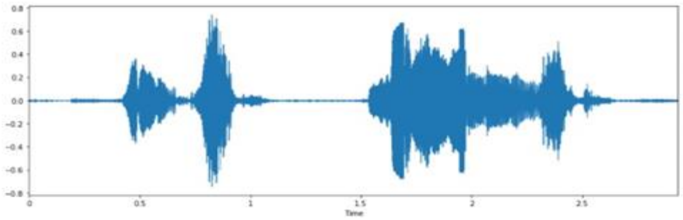


Fig.1. (d).Time Shifting

III. Convolution Neural Network:

CNN architecture we used for our experiment is shown in fig3. For the execution, we utilized Keras model-level library with the TensorFlow backend Programming of the architecture is done in Python using keras. It consists of 8 convolution filter, with ReLu activation, with a max-pooling layer and has 216 x 265 matrixes as input. The last stage comprises of a flattening and a dense layer of 192 neurons, followed by the emotion classifier. The distinctive feature of the CNN is the presence of sets of convolution and pooling layers. A convolution layer extracts the organized data with submatrices filters (strides) parsing on the two-dimensional information. A pooling layer sums up the output of the convolution network by conglomerating the values of the stride submatrix into a single value. The CNN architecture consists of many dense (fully connected) layers, and final layer is the classifier.

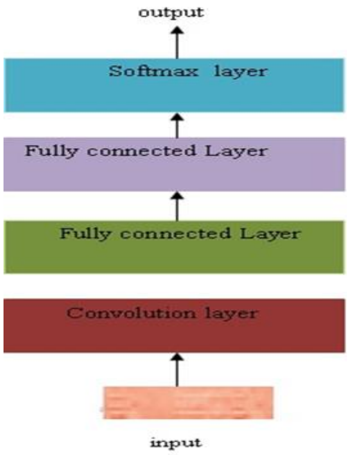


Fig.2 Representation of CNN architecture

IV. RESULT:

An accuracy of 46% was achieved in the model (without data augmentation) and an accuracy of 71 % is achieved using data augmentation. A 25% improvement is obtained by using combination of augmented and synthetic data.

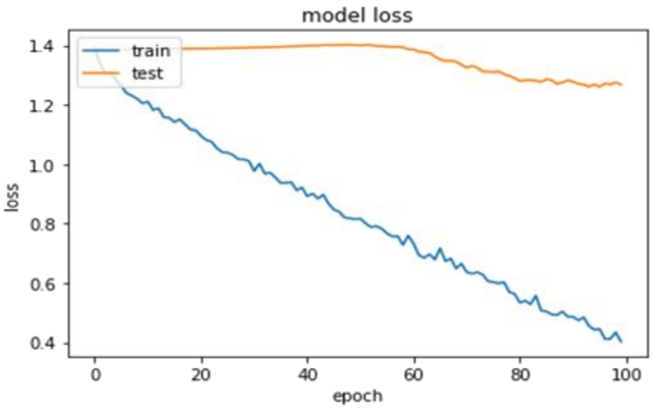


Fig.3.loss function without data augmentation

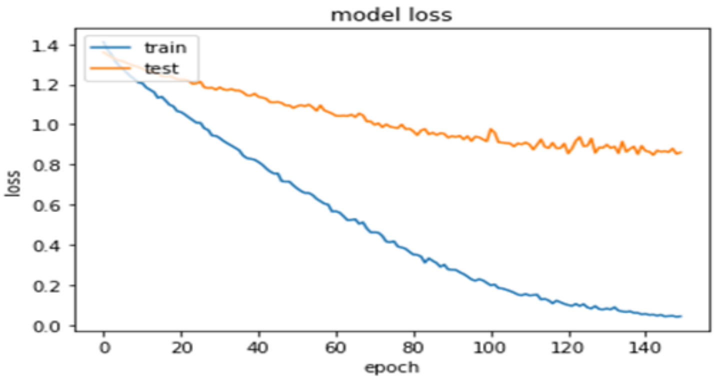


Fig.4. Loss function with dta Augmentation

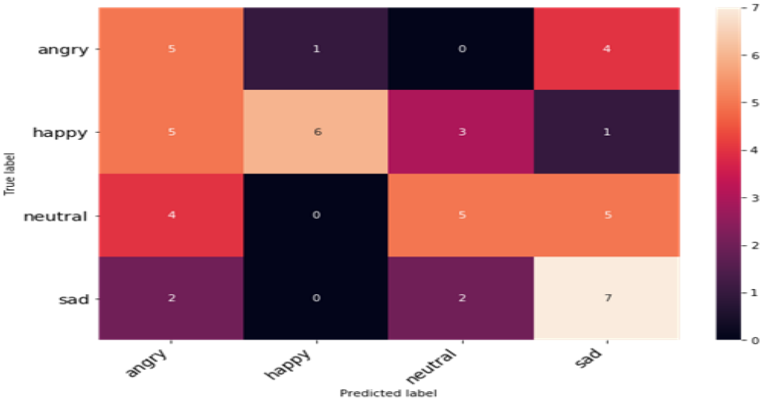
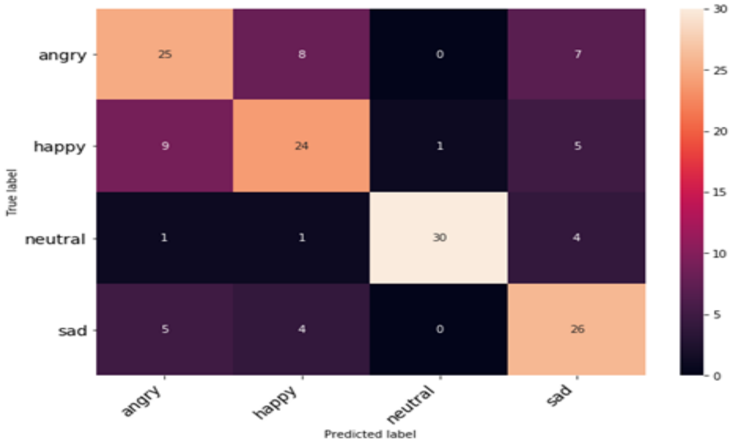


Fig.5 congusion Matrix Before Augmented data



True label \ Predicted label	angry	happy	neutral	sad
angry	25	8	0	7
happy	9	24	1	5
neutral	1	1	30	4
sad	5	4	0	26

Fig.6. Confusion Matrix after augmented Data

V. CONCLUSION :

It is observed that Data augmentation does help improve the accuracy slightly. We only introduced 2 augmentation methods. Perhaps if we include more it may make it more accurate. A 25% improvement is obtained by using combination of augmented and synthetic data. Next aim is to collect more real time data and increases our data base and further improves the model and implements it in real time data.

REFERENCES

[1] M. Reddy, "Depression: the disorder and the burden," Indian Journal of Psychological Medicine, vol. 32, no. 1, pp. 1, 2010.

[2] K. Huang,Chung H..W and Hsiang-chi.Fu “Mood detection from daily conversational speech using denoisingautoencoder and LSTM”, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5-9 March 2017

[3] Raman.G ,Hulyayalcin,“Improving Low Resource Turkish Speech Recognition with Data Augmentation and TTS” 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), 21-24 March 2019.

[4] Jisung Wang, Sangki Kim, Yeha Lee, “Speech Augmentation using Wavenet in Speech Recognition” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 12-17 May 2019.

[5] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation forspeech recognition,” Interspeech, 2015.

[6] G.R. Michael, Dr. Aditya Bihar Kandali,”Emotion recognition of Manipuri Speech using Convolution Neural Network”