Effects of Noise on RASTA-PLP and MFCC based Bangla ASR Using CNN

[1]Md. Raffael Maruf, [2]Md. Omar Faruque, [3]Salman Mahmood, [4]Nazmun Nahar Nelima, [5]Md. Golam Muhtasim, [6]Md.Jahedul Alam Pervez
Department of Electrical and Electronic Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
[1]raffaelmaruf94@gmail.com, [2]mohammadomarfaruque584@gmail.com, [3]s.m.shovon19@gmail.com,
[4]neelima.eee.173@gmail.com, [5]saminmuhtasim@gmail.com, [6]bdpervez12@gmail.com

Abstract—Though Bangla Automatic Speech Recognition (ASR) started its journey since a long time ago, a paltry amount of work is done on Convolutional Neural Network (CNN) based ASR. In this paper, we propose an ASR made with CNN where the performance of two feature extraction methods, namely Mel Frequency Cepstral Coefficients (MFCC) and Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) are compared on Bangla isolated words consisting of digits and speech commands. This paper contributes to the literature of Bangla ASR in three ways. Firstly, Effects of noise is experimented on Bangla speech commands as well as isolated words in CNN based ASR. Secondly, the performance of MFCC and RASTA-PLP are compared in noisy environment using CNN based classifier. Lastly, state-of-the-art accuracy is achieved in CNN based ASR which is 93.18%.

Keywords—Bangla ASR, CNN, MFCC, RASTA-PLP, DNN, ANN, Bangla Digits, Bangla Speech

I. INTRODUCTION

Automatic Speech Recognition (ASR) is defined as the process of converting, analyzing and recognizing particular patterns of speech signals employing algorithms in computational machines and used in a wide range of applications including health care, home automation, coping with disabilities, search engines etc. Despite the prospectus of ASR, Bangla, the 7th most spoken language with 250 million speakers around the world [1] lags behind others due to the lower number of conducted researches in this field. Thus, this paper is an attempt to fill up the void.

In this paper, we propose a convolutional neural network-based algorithm for Bangla speech recognition, which includes the most commonly used words and digits. For English, different feature extraction methods such as MFCC, LPC, LPCC, LSF, PLP, DWT etc. are used to filter out noise as well as identify words, while these are not much explored for Bangla. In this paper, we explored RASTA-PLP & MFCC methods to extract speech properties to analyze Bangla words. Furthermore, only a few pieces of research previously conducted on the Convolutional Neural Network (CNN) based Bangla word recognition system, despite it is one of the cutting edge neural networks used in AI. The effect of noise also has been examined here.

This paper is sectioned into the following parts: section II describes relevant literature, section III represents the methodology of the work, section IV gives the overview of the experimental setup, section V analyzes the result, conclusion and future work are discussed in section VI.

II. LITERATURE REVIEW

Convolutional neural network is known to be used for image classification but it can be effective too for speech recognition. Jui-Ting Huang et al. [2] have proved that CNN has an advantage over fully connected Deep Neural Network (DNN) by applying it on 1000 hours Kinect distant speech data. The CNN model with maxed out units gave 9.3%-word error rate reduction (WERR) over the same depth DNN with sigmoid units. A very deep CNN model is used on the Aurora4 dataset with adaptive noise, channel mismatch and AMI meeting corpus dataset [3]. The word error rate (WER) is 8.1% on the Aurora4 dataset which is closed to the accuracy of LSTM-RNN based model and on the AMI dataset, WER is 10% less than usual CNN model. In this paper [4], a performance comparison is shown among different feature extraction methods, i.e., MFCC, PNCC, PLP, RASTA-PLP on different isolated words in three different languages - Tamil, Bangla, Assamese. A prominent work has been done on Bangla digits ASR in [5]. CNN model-based ASR was implemented on Bangla short speech commands in [6]. The performance was compared among the three models. MFCC feature was extracted from the audio files in one approach, raw audio files were given as input in another model. Transfer learning was performed on another approach where a pre-trained model was used drilled on English short speech commands. Isolated Bangla words and corresponding speakers were detected in this [7] ASR system. A semantic modular time-delay neural network (MTDNN) was implemented as a classifier and fuzzy C means clustering technique was applied to increase the noise robustness.

III. METHODOLOGY

A. Dataset

The database used in this research work is a combination of Bangla digit sequences and short speech commands which are 11 unique words uttered by 120 different speakers containing male and female speakers between the ages of 7 and 60. The proposed CNN model was trained using 60% of the whole dataset. For validation, 20% of the dataset was allocated and rest for the test. For our experiment, we have recorded isolated spoken voices in room environment which can be regarded as almost noiseless. Then various urban noises were artificially added later on. This was done to explore the effects of various noises on the detection accuracy of various words. These noises are (1) Only Ac noise, (2) Only Cafe noise, (3) Only Library noise, (4) Only Rail station noise, (5) Only Street noise. The words we choose to train our model are shown in Table I.

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

TABLE I. ISOLATED SPEECH WORDS IN THE MENTIONED DATASET

Class label	Bengali words	English translation	Phonetic representation
0	এক	One	Ek
1	দুই	Two	Dui
2	তিন	Three	Tin
3	চার	Four	Char
4	পাঁচ	Five	Pach
5	শুরু	Start	Shuru
6	শেষ	End	Shesh
7	আসো	Come	Asho
8	যাও	Go	Jao
9	ডানে	Right	Dane
10	বামে	Left	Bame

As our actual data size is relatively small, we tried several augmentation techniques to overcome data scarcity by applying pitch shifting and time stretching on the training dataset.

B. *Theoretical Overview of CNN*

CNN [8] is a Deep Neural Network (DNN) architecture [9]. CNNs can discriminate spectral temporal patterns [10]. This is important in distinguishing between sounds. CNN comprises of numerous types of layers such as Convolutional Layer, Rectified Linear Units (ReLU), Pooling Layer, Dropout Layer, Fully Connected Layer.

- Convolutional Layer: Convolution layer performs convolution operation with the help of filters to extract features from data.
- Pooling Layer: The main purpose of the pooling layer is to reduce the special size of the extracted feature maps by deploying a window of arbitrary size called 'stride.'
- Rectified Linear Units (ReLU): It prevents the vanishing gradient problem occurring in 'sigmoids.' ReLU is easier to compute and generates sparsity [11].
- Dropout Layer: The dropout layer is used with other types of layers in a neural network. Dropout layers can be implemented with most types of layers such as convolutional layers, fully connected layers. It reduces over-fitting.
- Fully Connected Layer: It uses 'Softmax' function in the outer layer to make predictions. The main task of this function is to discriminate the input into various classes based on the training dataset.

IV. EXPERIMENTAL SETUP

A. *MFCC Model*

MFCC method is modeled after the human hearing. MFCC features were extracted from the audio files and fed to a CNN architecture. We will call this MFCC model.

B. *RASTA-PLP Model*

RASTA-PLP is a noise-robust feature extraction method. RASTA-PLP feature extraction method was deployed to extract features from audio files and then fed to the same CNN model. This is our RASTA-PLP model.

C. *CNN Architecture*

The proposed CNN model was built using 'Keras.' It was trained for 100 epochs while the batch size was 100. Dropout layer and l2 regularization were used to prevent overfitting. Max pooling layer was used to reduce feature size. The final layer was a SoftMax layer of 11 neurons. Grid search method was used for hyperparameter (i.e., dropout rate, batch size, neuron, learning rate, activation functions) optimization. Fig. 1 shows the dimensions of different layers in our proposed model.

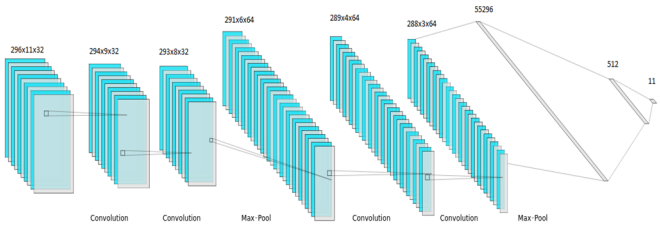

Fig. 1. The proposed CNN architecture.

V. RESULTS AND DISCUSSION

The CNN was trained using both MFCC features and RASTA-PLP features extracted directly from without and with augmentation datasets. 'LibROSA' python package was used for that. Data normalization had a negligible outcome on the result. Table II shows the comparison between MFCC model and RASTA-PLP model for both the datasets in terms of percentage accuracy. We can see that there is a substantial increase in accuracy after augmentation. We can also see that for the augmented dataset in room environment medium, RASTA-PLP performed better than MFCC model. Whereas in noisy environments, MFCC model performed better. Overfitting issue is more prominent in RASTA-PLP model than MFCC model. But the scenario completely reverses for dataset without augmentation. So our best accuracy comes when the data is augmented. From now, we will only discuss results trained on augmented data. Fig. 2 shows the epochs vs. accuracy graph of the MFCC model. Moreover, we have plotted the loss generated in every epoch in Fig. 3 for MFCC.

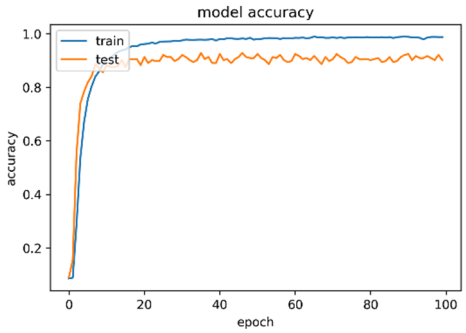

Fig. 2. Epoch vs. Accuracy for MFCC in room environment.

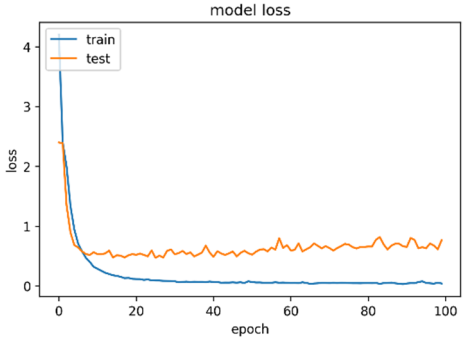

Fig. 3. Epoch vs. Loss for MFCC in room environment.

TABLE II. COMPARISON OF MODEL ACCURACY BETWEEN MFCC AND RASTA-PLP

Medium	Before augmentation				After augmentation			
	RASTA-PLP		*MFCC*		*RASTA-PLP*		*MFCC*	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
Room Environment	99.56	81.81	99.28	82.72	99.29	93.18	98.78	91.28
Ac noise	99.62	82.42	98.23	78.78	97.85	85.60	97.77	89.01
Cafe noise	98.55	81.51	98.36	78.78	98.82	87.87	98.28	90.15
Library noise	98.3	77.57	97.22	73.03	98.04	85.22	98.40	87.12
Rail station noise	97.6	77.27	87.12	66.66	96.69	82.19	96.10	90.53
Street noise	99.24	80.3	87.12	66.66	98.45	83.33	98.06	87.87

The accuracy and corresponding loss generated in every epoch of RASTA-PLP model were plotted in Fig. 4 and Fig. 5 respectively. Confusion matrix is a table layout which also enables us to visualize the performance of a neural network apart from model accuracy and loss curves. Fig. 6 shows the normalized confusion matrix of the MFCC model. From the confusion matrix, it appears that the model does exceedingly well in predicting certain words such as “Dui,” “Tin,” “Char,” “Pach,” “Shuru,” “Shesh,” “Asho,” “Jao” with a prediction accuracy which is well over 90%. Whereas the model does moderately well in predicting words such as “Ek,” “Bame,” “Dane.” Fig. 7 shows the normalized confusion matrix of the RASTA-PLP model. We can see that the model recognizes certain words such as “Pach,” “Shesh,” “Asho,” “Jao” with 100% accuracy, whereas the model has a tendency to label “Bame” as “Dane” as phonetically both words sound a bit similar.

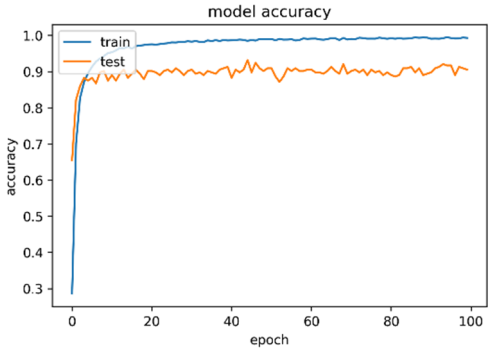

Fig. 4. Epoch vs. Accuracy for RASTA-PLP in room environment.

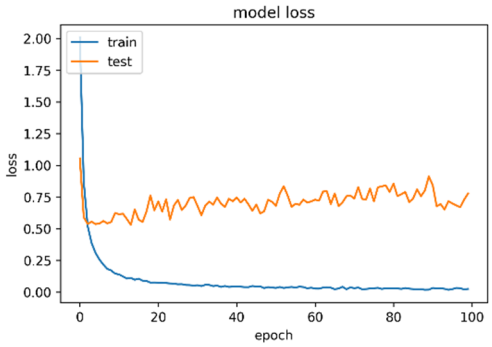

Fig. 5. Epoch vs. Loss for RASTA-PLP in room environment.

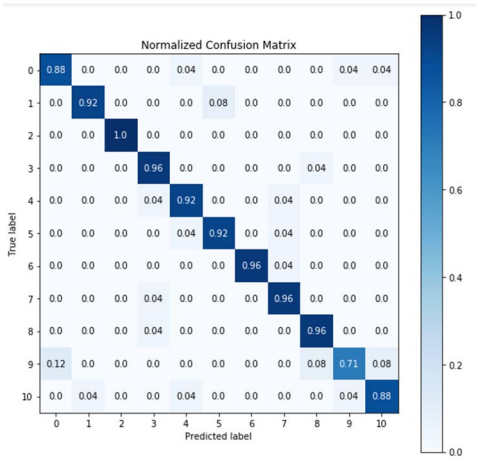

Fig. 6. Confusion Matrix of MFCC model in room environment.

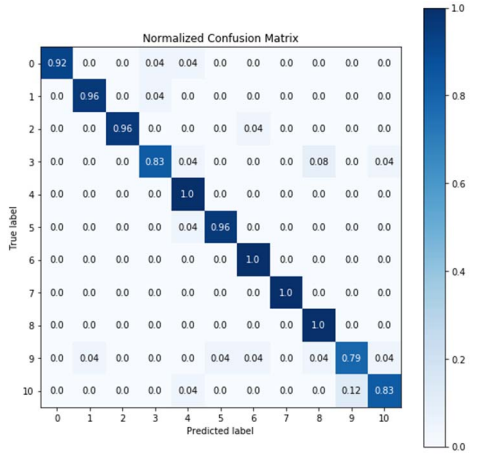

Fig. 7. Confusion Matrix of RASTA-PLP model in room environment.

Fig. 8 shows the comparison between MFCC model and RASTA-PLP model for both the datasets in terms of percentage accuracy for Training and Testing cases. Fig. 9 shows the comparison of prediction accuracy of various words between RASTA-PLP and MFCC models on various mediums in terms of percentage. For instance, in Café noise medium, the prediction accuracy for the word “Bame” was 83% and 96% for RASTA-PLP and MFCC model respectively. Likewise, for all the rest of the words.

We also compared our best model with contemporary work. Table III shows the performance comparison of our best model with contemporary models.

TABLE III. COMPARISON OF ACCURACY BETWEEN PROPOSED CNN BASED MODEL WITH CONTEMPORARY CNN BASED MODEL

Model	Train accuracy	Test accuracy
Proposed RASTA-PLP model (Room Environment)	99.29%	93.18%
Reference MFCC model [6]	85.44%	74.01%

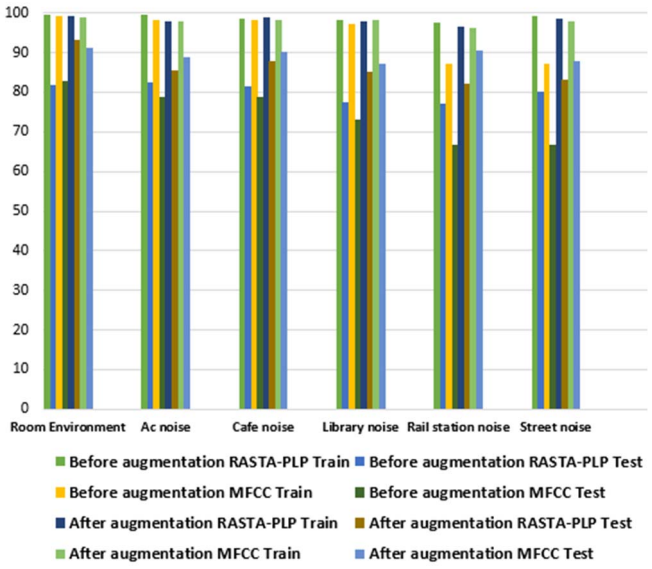

Fig. 8. Comparison of model accuracy between MFCC and RASTA-PLP.

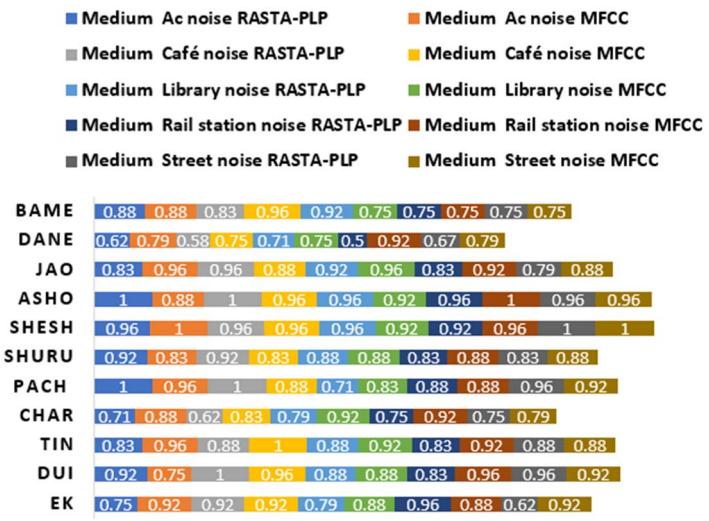

Fig. 9. Comparison of prediction accuracy of various words between RASTA-PLP and MFCC models on various mediums.

VI. Conclusion And Future Work

In this paper, performance of MFCC and RASTA-PLP are explored on Bangla isolated word recognition using CNN based ASR in room environment and noisy medium. Experimental results show that RASTA-PLP outperforms MFCC in room environment but MFCC performs better in noisy environment. In the future, we would like to increase the vocabulary range to implement the ASR model in hardware systems, for instance, Bangla voice-controlled wheelchairs as well as robot arms. Hopefully, this effort will bridge the gap between technology and the Bangla speaking population.

VII. References

[1] D. Eberhard, G. Simons, and C. Fennig, "Ethnologue: Languages of Asia," 2019.

[2] J. T. Huang, J. Li, and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Aug. 2015, vol. 2015-Augus, pp. 4989–4993, doi: 10.1109/ICASSP.2015.7178920.

[3] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 24, no. 12, pp. 2263–2276, Dec. 2016, doi: 10.1109/TASLP.2016.2602884.

[4] R. Sriranjani, B. M. Karthick, and S. Umesh, "Experiments on front-end techniques and segmentation model for robust Indian Language speech recognizer," in 2014 20th National Conference on Communications, NCC 2014, 2014, doi: 10.1109/NCC.2014.6811284.

[5] G. Muhammad, Y. A. Alotaibi, and M. N. Huda, "Automatic speech recognition for Bangla digits," in ICCIT 2009 - Proceedings of 2009 12th International Conference on Computer and Information Technology, 2009, pp. 379–383, doi: 10.1109/ICCIT.2009.5407267.

[6] S. Ahmed Sumon, J. Chowdhury, S. Debnath, N. Mohammed, and S. Momen, "Bangla Short Speech Commands Recognition Using Convolutional Neural Networks," in 2018 International Conference on Bangla Speech and Language Processing, ICBSLP 2018, Nov. 2018, doi: 10.1109/ICBSLP.2018.8554395.

[7] M. Y. A. Khan, S. M. M. Hossain, and M. M. Hoque, "Isolated Bangla word recognition and speaker detection by semantic modular time delay neural network (MTDNN)," in 2015 18th International Conference on Computer and Information Technology, ICCIT 2015, Jun. 2016, pp. 560–565, doi: 10.1109/ICCITechn.2015.7488134.

[8] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 2018–2025, doi: 10.1109/ICCV.2011.6126474.

[9] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Oct. 2015, vol. 07-12-June, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[10] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Process. Lett., vol. 24, no. 3, pp. 279–283, Mar. 2017, doi: 10.1109/LSP.2017.2657381.

[11] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Journal of Machine Learning Research, 2011.