Comparative Study of Effective Augmentation Method for Bangla ASR Using Convolutional Neural Network

Md. Raffael Maruf, Md. Omar Faruque, Md. Golam Muhtasim, Nazmun Nahar Nelima, Salman Mahmood, and Md. Maiun Uddin Riad

Abstract Data scarcity is the main obstacle to get top-notch accuracy in neural network-based automatic speech recognition (ASR). To solve this problem, using data augmentation is a very familiar phenomenon nowadays. But in Bangla ASR, this technique is seldom used. This paper illustrates the benefits of data augmentation. In this work, the most effective augmentation methods have been explored for Convolutional Neural Network (CNN)-based Bangla word recognition system. A few types of augmentation methods have been implemented in this task, e.g., time stretching, background noise injection and pitch shifting. This experiment is performed on Linear Prediction Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) feature extraction methods for comparative analysis. In LPCC extraction, slow-down augmentation is the most efficient one, whereas, in MFCC, positive pitch shifting augmentation is the hugely successful one at raising the accuracy rate. Overall, time-stretching is the most effective augmentation method that has consistently given better accuracy in both features. Contrariwise, noise injection is a less effective method in both cases. The consequence of using all these augmentation techniques is the escalation of accuracy up to 27.27%. To the best

Md. R. Maruf · Md. O. Faruque (✉) · Md. G. Muhtasim · N. N. Nelima · S. Mahmood · Md. M. U. Riad
Department of Electrical & Electronic Engineering, Ahsanullah University of Science and Technology, Dhaka 1208, Bangladesh
e-mail: mohammadomarfaruque584@gmail.com

Md. R. Maruf
e-mail: raffaelmaruf94@gmail.com

Md. G. Muhtasim
e-mail: saminmuhtasim@gmail.com

N. N. Nelima
e-mail: neelima.eee.173@gmail.com

S. Mahmood
e-mail: s.m.shovon19@gmail.com

Md. M. U. Riad
e-mail: maiunuddinriad@gmail.com

A. E. Hassanien et al. (eds.), *Emerging Technologies in Data Mining and Information Security*, Advances in Intelligent Systems and Computing 1286,
https://doi.org/10.1007/978-981-15-9927-9_20

of our knowledge, this research paper is a pioneer work to inquire about the most effective augmentation method for Bangla ASR.

Keywords Convolutional neural network · Bangla word recognition system · Automatic speech recognition

1 Introduction

Automatic speech recognition refers to a technology to utilize the voice with the computer and resemble the conversation. ASR technologies nowadays emphasize using voice as an interface for workplace automation hubs like lighting, audio-visual equipment, camera, security system, etc.

Real-time ASR and AI systems perform to monitor trading room conversations and other financial advice conversations. ASR technology is exposing speech pathology system where patients can utilize the speech rehabilitation programs. Voiceprint biometric system becomes stabilized because of the speaker recognition procedure. While Bangla is the seventh most widely spoken language with 250 million speakers worldwide, the development of the Bangla ASR system trails behind others due to the limited number of research works conducted in this area, and this work seeks to fill in the gap.

This paper proposes a convolutional neural network (CNN)-based speech recognizer for Bangla frequently used command words and digits. But it requires an enormous amount of data to train a convolutional neural network to achieve significant accuracy. The solution to this problem is to collect more data or manipulating the existing data to increase the size of the dataset. In this paper, we explored the effective audio data augmentation methods for MFCC and LPCC features to maximize prediction accuracy.

2 Literature Review

Hybrid Deep Neural Network (DNN)—Hidden Markov Model (HMM)-based speech recognition system showed high accuracy. But CNN-based ASR outperformed DNN by reducing the error rate up to 10% in this paper [1]. Nithya Davis et al. [2] performed different data augmentation techniques on the same dataset, where it was found that LPCC-based augmentation had better performance than other augmentation methods. With data augmentation, it was found that it may increase accuracy by more than 5%. In this paper [3], 'SpecAugment' augmentation method was applied to the training dataset of a speech recognition system, where it had converted the ASR system from an over-fitting to an under-fitting problem. The achieved performance was 6.8% WER in LibriSpeech on test other without the use of a language model. Reference [4] worked on LPCC- and MFCC feature-based isolated Bangla word recognition using several classifiers, i.e., nearest neighbor (DTW), neural net (NN), word-based HMM (HMMw) and phoneme-based HMM (HMMp). They generated synthetic samples from real samples. They reported a significant increase in isolated word detection rate in all classifiers trained with both the samples compared to training with just the real samples. In this paper [5], a performance comparison between three different CNN models was shown for Bangla short speech command. The dataset was used with MFCC feature extraction in one model and raw in the other and the third approach used transfer learning where the MFCC model showed better performance than the others. Reference [6] presented MFCC feature-based Bangla speech recognition system using convolution neural network. They demonstrated an increase in classifier performance by varying dropout rate and using data augmentation techniques. They reported an accuracy of 86.058% for isolated speech to text conversion in the CNN model. An acoustic model for Bangla digit recognition was proposed [7] using DBN, which is a probabilistic generative ANN with feature detectors comprise of multiple layers. Comparing with the other prominent methods, the proposed model showed satisfactory accuracy and outnumbered the rest.

3 Methodology

3.1 Data Collection

The dataset used in this research work consists of Bangla digit sequences and short speech commands. Eleven unique words uttered by 306 different speakers were recorded in a room environment containing both male and female speakers between the ages of 7 and 60 with a duration of 3 s. 60% of the whole dataset was used for training purposes. 20% of the dataset was allocated for validation and the rest for the test. Table 1 shows the isolated speeches that comprise the used dataset.

3.2 Data Augmentation

Syntactic data for audio can be generated by applying various techniques. The techniques that have been used in the proposed work are listed below.

1. Time stretching: Time-stretch an audio series by a ‘Stretch factor’ while keeping the pitch unchanged. If factor >1, then the signal speeds up. If factor <1, then the signal is slowed down. In this work, each audio sample was time-stretched by 6 factors: {1.33, 1.5, 2, 0.5, 0.66, 0.75}.
2. Pitch shifting: Shift the pitch of an audio signal by ‘Semitone factor’. For positive pitch shifting that resembles male voice, factor >0. For negative shifting that

Table 1 Isolated speech words in the dataset

Bengali words	English translation	Phonetic representation
এক	One	Ek
দুই	Two	Dui
তিন	Three	Tin
চার	Four	Char
পাঁচ	Five	Pach
শুরু	Start	Shuru
শেষ	End	Shesh
আসো	Come	Asho
যাও	Go	Jao
ডানে	Right	Dane
বামে	Left	Bame

simulates female voice, factor <0. In this paper, each audio sample was pitch-shifted by 6 factors: {2, 3, 4, −2, −3, −4}.

3. Noise injection: In this technique, an audio sample is mixed with other records consisting of various environmental background noises. Each sample was mixed with three environmental noises: {Washing dish, Running tap, White noise}.

3.3 CNN Architecture

Convolutional Neural Network (CNN) spreads its footmark on the object, faces and sound recognition [8]. This consists of neurons that contain weights and biases. Neurons take some data for the beginning but it circulates the procedure and ends up with the non-linearity. The architecture of CNN connects with the human brain neuron pattern and the Visual cortex inspires them. Computer-based deep learning produces an algorithm where it is perfected with time [9].

The architecture of CNN consists of several layers; there will be input and output layer and in the mid other layers take part in the operation. Those are called the hidden layer. The first hidden layer is called convolutional layer where it convolves with the dot product. After that, the RELU layer plays a role then follows the pooling

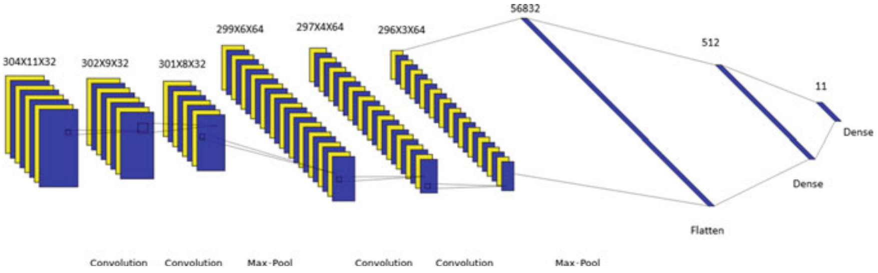

Fig. 1 The proposed CNN architecture

layer where the data shrinks. Finally, the fully connected layers and normalization layers end up the hidden portion and address the data to an output.

The proposed CNN model was built using ‘Keras’, which is an open-source neural network library written in Python. The model was trained for 100 epochs and the batch size was 100. Dropout layer and l2 regularizations were used to prevent overgeneralization of the training dataset. The value of lambda (regularization hyperparameter) was 0.001. We used 25% dropout to reduce overfitting. The proposed model in this paper uses ‘ADAM’ optimizer with a learning rate of 0.001. Max pooling layer was used to reduce feature size. SoftMax activation function was used in the final layer. Figure 1 shows the dimensions of different layers in our proposed model.

3.4 Linear Predictive Cepstral Coefficients (LPCC)

This LPCC is a derivative of linear predictive coding (LPC). Steps for deriving LPCC are given below in chronological order.

- Pre-emphasis: The sample speech has to be passed through a filter. The goal is to flatten the signal and to make it less prone to noise [10].
- Framing: The resulting speech is divided into time frames.
- Windowing: Frames are multiplied by a hamming window to reduce edge effect.
- Calculate the LPC: Auto-correlation is applied through the sample. Using vector quantization method in all frames, LPC coefficients have to be extracted for each sample.
- Calculate LPCC: In the final step, LPCC features are calculated from LPC features.

3.5 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) is a popular feature extraction method. Steps for deriving MFCC are given below in chronological order.

- Pre-emphasis: To amplify the high frequency, we have to apply pre-emphasis filter
- Framing: After the pre-emphasis, we need to split the signal into short-time frames.
- Hamming window: To reduce discontinuity, each frame is infiltrated with Hamming Window.
- Fast-Fourier transform: It is applied on time frames for frequency–domain conversion.
- Mel-Scale filter bank: Mel-Scale filter bank is used to convert the spectrum into the Mel scale, which simulates human hearing.
- Discrete cosine transform: It is applied to transform the log Mel spectrum into the time domain.
- Delta energy and delta spectrum: Appending time derivatives on static MFCC will give better performance [11].

4 Results and Discussion

The neural network was trained for 100 epochs using both MFCC and LPCC features extracted from without augmentation dataset and with augmentation dataset. Test accuracies of MFCC and LPCC models for each augmentation case are presented in Fig. 2. Before augmentation, test accuracies for LPCC and MFCC models were 80.27% and 72.13% respectively.

4.1 Effect of Augmentation in LPCC Model

The percentage increase in test accuracy for various augmentation methods is presented in Table 2. From Table 2, if we compare the test accuracies between speed up case (all factors applied, i.e., 1.33, 1.5, 2) and speed down case (all factors applied, i.e., 0.75, 0.66, 0.5) then it is apparent that slow down was more effective. Likewise, negative pitch method produced better results than positive pitch shift method. Among all the noise injection methods, data augmentation by adding white noise is the most effective one. In order of effectiveness, slow down was most effective then speed up and then negative pitch shift and noise injection. Positive pitch shift was least effective. When all these methods are applied together, test accuracy was 89.28%, which is an 11.22% increase.

4.2 Effect of Augmentation in MFCC Model

If the test accuracies between speed up method (all factors applied) and slow down method (all factors applied) are compared, then it is apparent that speed up was more effective in MFCC model. Similarly, positive pitch shift method produced better

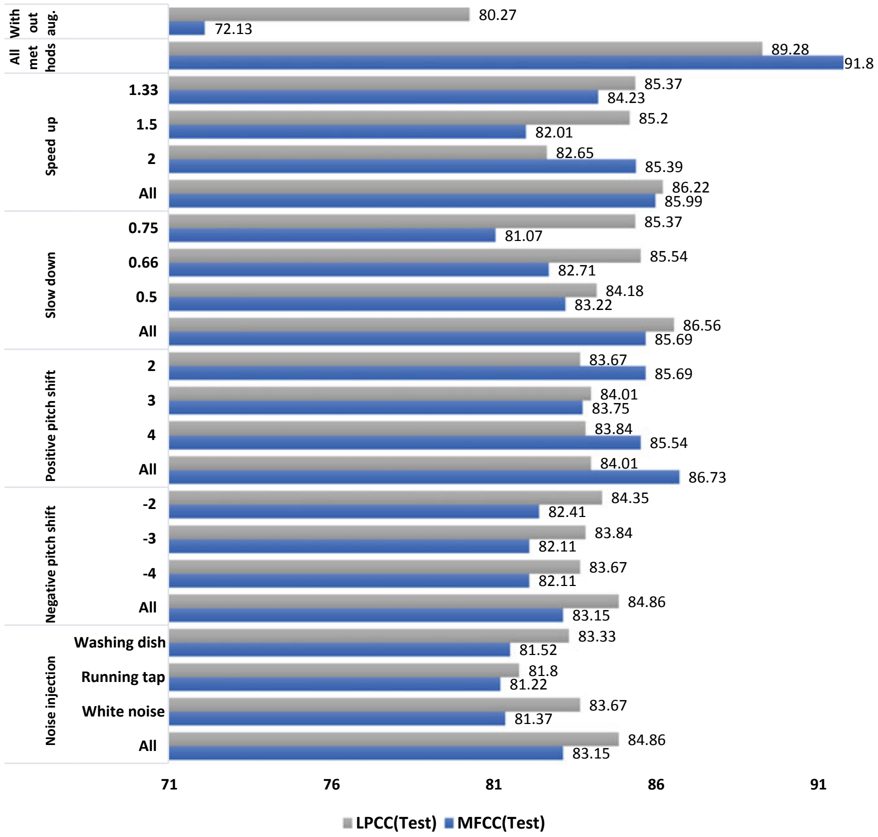

Fig. 2 Comparison of accuracies of MFCC and LPCC models for individual augmentation cases

Table 2 Comparison of increase in accuracy for LPCC and MFCC models

Model	Speed up (All) (%)	Slow down (All) (%)	Positive pitch shift (All) (%)	Negative pitch shift (All) (%)	Noise injection (All) (%)	All methods (%)
LPCC	7.41	7.84	4.66	5.72	5.72	11.22
MFCC	19.21	18.80	20.24	15.28	15.28	27.27
Average of LPCC and MFCC	13.31	13.32	12.45	10.50	10.50	19.25

results than negative pitch shift method. Among all the noise injection methods, data augmentation by adding washing dish as background noise is the most effective one. In order of effectiveness, positive pitch shifting was most effective then speed up and then slow down. Negative pitch shift and noise injection were least effective. When

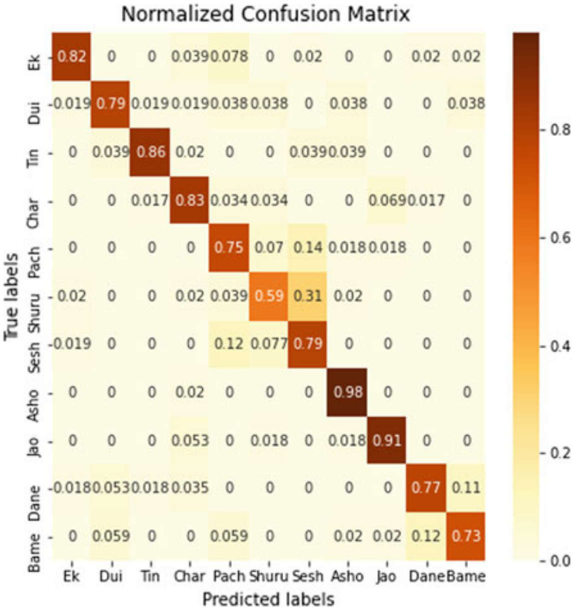

Fig. 3 Confusion matrix of LPCC model before augmentation

all these methods are applied together, test accuracy was 91.8%, which is a 27.27% increase.

4.3 Overall Evaluation

The effect of augmentation is more prominent in the MFCC model since it outperformed the LPCC model on augmented dataset although LPCC performed better on without augmentation dataset. On average, augmentation using time stretching method was found to be the most effective and noise injection was the least effective for both models. Adding white noise proved to be most effective for both MFCC and LPCC models among all the noise injection methods. Whereas running tap noise was the least effective. Figure 3, Fig. 4, Fig. 5, and Fig. 6 show the impact of data augmentation (all methods) on prediction accuracy per class for the LPCC model and MFCC model, respectively.

5 Conclusion

This research paper has proposed the most-effective augmentation methods for increasing test accuracy for MFCC- and LPCC feature-based Bangla ASR using CNN. In LPCC extraction, slow-down augmentation is the most efficient one,

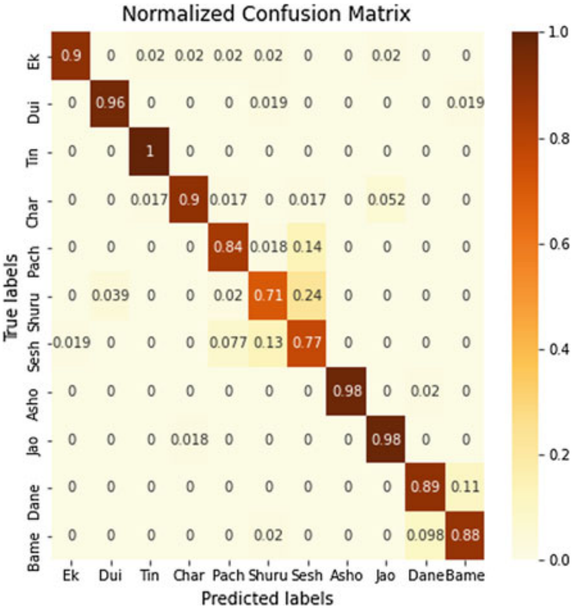

Fig. 4 Confusion matrix of LPCC model after augmentation

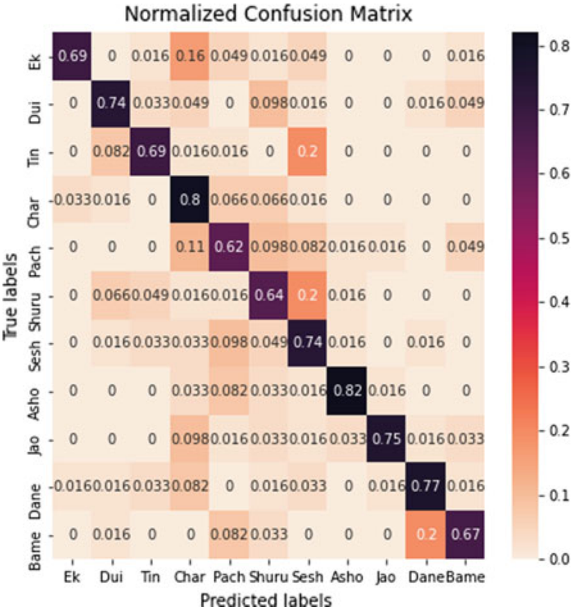

Fig. 5 Confusion matrix of MFCC model before augmentation

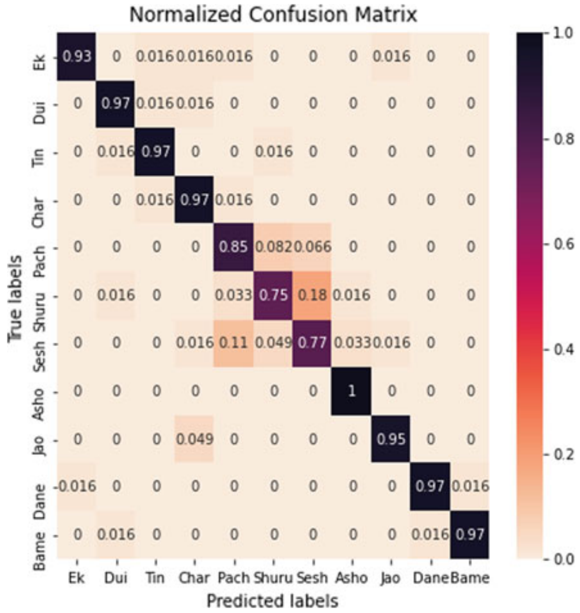

Fig. 6 Confusion matrix of MFCC model after augmentation

whereas, in MFCC, positive pitch shifting is the hugely successful one at raising the accuracy rate. Based on average increment in accuracy for both MFCC and LPCC models, slow down and speed up both produced almost similar results and they were most effective. Positive pitch-shifting method produced the second-best results. Negative pitch shifting and background noise injection generated identical increments and proved to be the least effective. In conclusion, it can be said that to get optimal performance from a CNN-based classifier for automatic speech recognition; data augmentation using the time stretching method, namely, speed up and slow down, needs to be prioritized.

References

1. Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE Trans. Audio Speech Lang. Process. **22**, 1533–1545 (2014). https://doi.org/10.1109/TASLP.2014.2339736
2. Davis, N., Suresh, K.: Environmental sound classification using deep convolutional neural networks and data augmentation. In: 2018 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2018, pp. 41–45. Institute of Electrical and Electronics Engineers Inc. (2019)
3. Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q. V.: SpecAugment: A simple data augmentation method for automatic speech recognition. In: Proceedings of Annual Conference on International Speech Communication Association INTERSPEECH, 2019-September, pp. 2613–2617 (2019). https://doi.org/10.21437/Interspeech.2019-2680

4. Chakraborty, R., Garain, U.: Role of synthetically generated samples on speech recognition in a resource-scarce language. In: Proceedings—International Conference on Pattern Recognition, pp. 1618–1621 (2010)
5. Ahmed Sumon, S., Chowdhury, J., Debnath, S., Mohammed, N., Momen, S.: Bangla short speech commands recognition using convolutional neural networks. In: 2018 International Conference on Bangla Speech and Language Processing, ICBSLP 2018, Institute of Electrical and Electronics Engineers Inc. (2018)
6. Islam, J., Mubassira, M., Islam, M.R., Das, A.K.: A speech recognition system for Bengali language using recurrent neural network. In: 2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019, pp. 73–76.Institute of Electrical and Electronics Engineers Inc. (2019)
7. Ahmed, M., Shill, P.C., Islam, K., Mollah, M.A.S., Akhand, M.A.H.: Acoustic modeling using deep belief network for Bangla speech recognition. In: 2015 18th International Conference on Computer and Information Technology, ICCIT 2015, pp. 306–311. Institute of Electrical and Electronics Engineers Inc. (2016)
8. Prabhu, R.: Understanding of Convolutional Neural Network (CNN)—deep learning. https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148 (2020). Accessed 29 Apr 2020
9. Saha, S.: A Comprehensive Guide to Convolutional Neural Networks —the ELI5 way. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (2020). Accessed 29 Apr 2020
10. Sandanalakshmi, R., Martina Monfort, V., Nandhini, G.: A novel speech to text converter system for mobile applications. In. Int. J. Comput. Appl. **73**, 7–13 (2013). https://doi.org/10.5120/12991-9886
11. Das, P.P., Allayear, S.M., Amin, R., Rahman, Z.: Bangladeshi dialect recognition using Mel Frequency Cepstral Coefficient, Delta, Delta-delta and Gaussian Mixture Model. In: Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016, pp. 359–364. Institute of Electrical and Electronics Engineers Inc. (2016)