2019 IEEE 4th International Conference on Computer and Communication Systems

A Speech Recognition System for Bengali Language using Recurrent Neural Network

Jahirul Islam
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
e-mail: jahir.ewubd@gmail.com

Masiath Mubassira
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
e-mail: masiathmubassira@gmail.com

Md. Rakibul Islam
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
e-mail: rakibrk.ewu@gmail.com

Amit Kumar Das
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
e-mail: amit.csedu@gmail.com

***Abstract*—Speech recognition is the most interactive technology between a human and a machine. Over the past 70 years, tremendous work has been accomplished in this fundamental area of speech communication. However, implementation of the Bengali language is unsubstantial in the field of Human-Computer Interaction. This research paper tried to implement convolution neural network technique for creating a speech recognition system in the Bengali language. We also implemented recurrent neural network to find the Bengali character probabilities which were then improved further by using CTC loss function and language model. This paper implemented Bengali language consisting of diacritic characters and such languages are very much difficult to train in a model.**

Keywords-Speech recognition, MFCC, CNN, RNN, CTC, Language model.

I. INTRODUCTION

Speech recognition system is widely used since it provides dictation ability and people can use this technology to control devices very easily. It is assumed that 50% of all searches will be voice searches by 2020 [1]. At the beginning of the twenty-first century, there has been widespread use of speech recognition. It is useful for persons who find typing difficult and also helps those with spelling difficulties, including users with dyslexia [2].

Speech recognition system has been developed into many languages such as English, German, and Mandarin because of the high demand of these languages. Development of speech recognition system for Bengali did not attract much attention around the globe. Around 245 million people have their native language as Bengali. About one-sixth population of the world speaks in Bengali [3]. It is ranked 7th based on the number of speakers [4]. So, our research focused primarily on building an efficient model for speech to text synthesis for the Bengali language.

This paper proposed two models for Bengali speech recognition systems. The first model used Convolutional Neural Network (CNN) and wanted to find whether we could successfully recognize spoken Bengali language. Our second Bengali speech system used recurrent neural network as the deep learning technique. The predicted output had been merged into sentences by using Connectionist Temporal Classification (CTC). The transcription was then improved by using language model.

The limitations of the first model were the lack of recorded speech data and text data. The challenge for the second model was it needed a powerful tool to train itself with thousands of data. Also, the Bengali language has diacritic characters which are very tough to implement in RNN model compared to other languages such as English.

This paper had the following contribution:

- Our proposed system can work offline.
- This paper did not need to identify phonemes and very less preprocessing was needed.
- It is an open sourced system.
- It could be integrated with other systems.

The rest of the paper is organized as follows. The Section II describes some of the state-of-the-art works related to Bengali speech recognition system Later, it provides the proposed work for CNN system in Section III. Section IV contains the architecture of RNN system. The results and analysis portion are contained in Section V, and finally, Section VI is the conclusion.

II. RELATED WORK

Forty-seven phonemes were selected for the Bengali language. Features were extracted from these frames by using MFCC. HTK and CMU SPHINX toolkit had been used for the acoustic model. Their model worked well with data from the young population, but the accuracy rate decreased with data from the old population [5].

978-1-7281-1322-7/19/$31.00 ©2019 IEEE

Firstly, the speech was recorded, and the noise was eliminated by using an adaptive filter. MFCC was used to extract the features from the signal. Then phoneme based and word-based HMM models had been used for training the continuous and isolated-word based speech recognition systems. The isolated-word recognizer could only recognize words that were in the dictionary and were spoken by the same speaker and the mood of the speaker had to be the same [6].

This paper used four speech recognition models and compared the recognition rate and elapsed time of each model. The first model used MFCC and Dynamic Time Warping. The second model used Linear Predictive Coding and Dynamic Time Warping. The third model used MFCC, Gaussian Mixture and posterior probability function and this had the highest accuracy of 84%. The fourth model used MFCC and Linear Predictive Coding and Dynamic Time Warping [7].

This paper compared monophone based acoustic model with a triphone based acoustic model in Bengali. Triphone based acoustic model needed far more training data than monophone based acoustic model. So triphones were clustered according to the similarities on phoneme. The results showed that tied-state triphone model performed better than the monophone based model [8].

This paper recorded 10 Bengali digits from 10 speakers. The features were extracted by using MFCC. These features of 5 speakers were used to train the model by using backpropagation neural network. The digits from 0 to 9 for the rest of the five speakers were used to test the model. This model achieved a recognition rate of about 96.332% for known speakers (speaker dependent) and 92% for unknown speakers (speaker independent) [9].

III. Architecture of CNN System

In our recognition system, preprocessing, feature extraction and recognition were three important aspects. They are as follows:

A. Preprocessing

First of all, we recorded voice data with a microphone. We needed to record data in wav format of 16 bits and frequency of 16000 Hz. This wav data was converted into numpy array because wav file needed huge memory capacity and its processing time was high.

B. Feature Extraction

After getting the preprocessed data, we extracted the features by using mel-frequency cepstral coefficients (MFCC). We read voice file from the specific directory and then computed MFCC from that given voice file. MFCC vectors might vary in size for different audio input, but our model could not incorporate with files of different size. So, we specified the duration of every voice sample as 1 second.

C. Recognition Model

We created a recognition model using convolutional neural network (CNN). We implemented neural network containing 7 levels where the first layer was the input layer, the last layer was the output layer, and in between, there were five hidden layers. We fixed the dropout rate as 0.25, so that every layer reduces the weight of the previous layer at a rate of 0.25.

IV. Architecture of RNN System

This model used recurrent neural network to train this architecture. The output of the neural network was a matrix consisting of character probabilities over time. Connectionist Temporal Classification (CTC) loss function had been used to maximize the probability of the correct transcription being predicted. For improving the text sequence further, language model has been implemented in this paper.

A. Dataset Used

On full dataset, there were in total 508 speakers. This dataset contained in total 200,000 wav audio files approximately, but only 33000 audio files were used for our task. Among these 33000 data, 80% have been used as training data, 90% as validation data and 10% as test data. The total duration of 33000 datasets was 33 hours approximately. Dataset content of audio was converted into wav files. The sampling rate was 16000, and one channel was used. The text file was 16-bit PCM encoded transcripts. All the text data of the wav files were in Bengali.

The dataset contained diacritic characters. Thus, it is more difficult to break down Bengali audio files into audio transcription since it contains diacritic characters.

B. Formulation of RNN

The neural network had five layers. The input was given into three fully connected layers, followed by a bidirectional RNN layer, and finally another fully connected layer. Let a single utterance be a and the label be b for a training set,

$$Z = \{(a^{(1)}, b^{(1)}), (a^{(2)}, b^{(2)}), ...\} \quad (1)$$

Every utterance, $a^{(i)}$, consisted of a time-series of length $L^{(i)}$ where each time slice consisted of a vector of audio features, $a_t^{(i)}t=1, 2, ..., L^{(i)}$.

At time t, for the first layer, the output depended on the spectrogram frame $a^{(t)}$ and also on the preceding and foremost frames of $a^{(t)}$. Thus, for every time t, the first three layers were computed by:

$$l_t^{(k)} = g(X^{(k)} l_t^{(k-1)} + b^{(k)}) \quad (2)$$

where g(z) = min {max {0, z},20} was the clipped rectified-linear (ReLU) activation function and W(k), b(k) were the weight matrix and bias parameters for layer k respectively.

The next layer was a bidirectional RNN layer. This layer consisted of two sets of hidden units: a set which contained forward recurrent network, l(f), and a set which contained backward recurrent network l(b):

$$l_t^{(f)} = z\,(X^{(4)} lt^{(3)} + X_r^{(f)} l_{t-1}^{(f)} + b^{(4)}) \quad (3)$$

$$l_t^{(b)} = z\,(X^{(4)} lt^{(3)} + X_r^{(b)} l_{t+1}^{(b)} + b^{(4)}) \quad (4)$$

It must be noted that $l^{(f)}$ should be computed sequentially from t = 1 to t = T(i) for the i'th utterance, whereas the units $l^{(b)}$ have to be computed from t=T(i) to t=1 sequentially in reverse order. The 5th layer took input from both the forward and backward units. The output consisted of the softmax

function that provided output as character probabilities for each character in the alphabet.

C. *Improving Accuracy with CTC and Language Model*

Once this paper had computed a prediction for the characters in the audio transcription, we used CTC to find the transcription.

We then tried to find a sequence s that would maximize the combined objective:

$$R(s) = \log (P(s|x)) + \alpha \log (P_{lm}(s)) + \beta \text{ word count}(s) \quad (5)$$

where α and β were tunable parameters which were set by cross-validation, and it controlled the trade-off between the language model and the recurrent neural network. The term P_{lm} denoted the probability of the audio transcription [10].

V. Results and Analysis

A. *The Following Analysis Is Based on the first Model Which was Based upon CNN System.*

We took 30 samples of each word to train our model. We did not add any noise to our sample. We trained our model with the raw data. Then for speech recognition, we took a vocabulary of 100 words and the length of the file was 100 seconds. Those 100 words were the combination of the trained word but with a different sample. Then we used our model to recognize the speech.

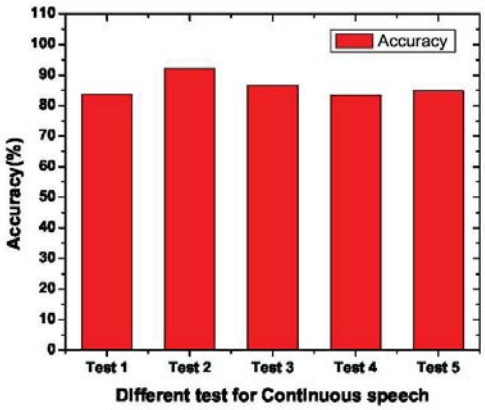

Figure 1. Accuracy in the different test for continuous speech (speaker dependent)

Recognizing continuous speech with Bengali Corpus had an average accuracy rate of 74% [5]. Recognition system using feature extraction and spectrum had the maximum accuracy rate of 84% [7]. From figure 1, we can see that the recognizer presented in this paper had an average accuracy rate of 86.058%.

B. *The Following Analysis Is Based on the second Model Which was Based upon the RNN System.*

The comparisons were done by using dropout values. The hyperparameters were changed and the new values were recorded. The hidden cells in RNN were 2048. The learning rate was 0.0001. For each experiment, a total of 50 epochs were used.

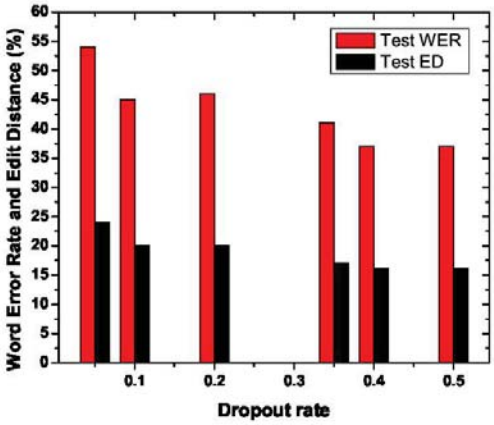

Figure 2. Word Error Rate (WER) and Edit Distance (ED) change with respect to dropout

From figure 2, we could see that as the value of dropout increased, the value of the test WER decreased. This indicated that as we discarded more parameters by using dropout, the model could not memorize the outcomes. Thus, the model can predict better outcome by using the test data.

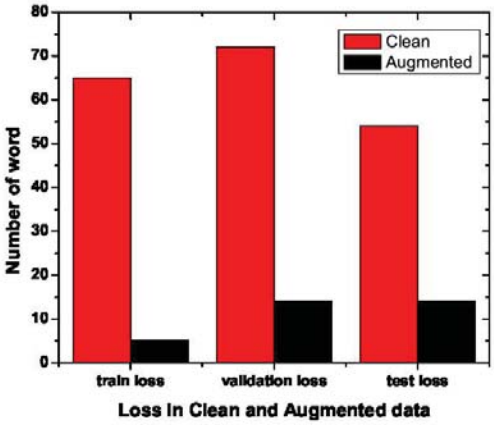

Figure 3. Loss comparison between clean and augmented dataset

The hidden cells used for figure 3 were 1024. Train loss, deviation loss, and test loss were more for clean data compared to augmented data. This showed that if we

augmented the data, that is, if we changed the pitch, speed and other characteristics of the audio data, then our model could provide a better-predicted outcome of the audio transcription.

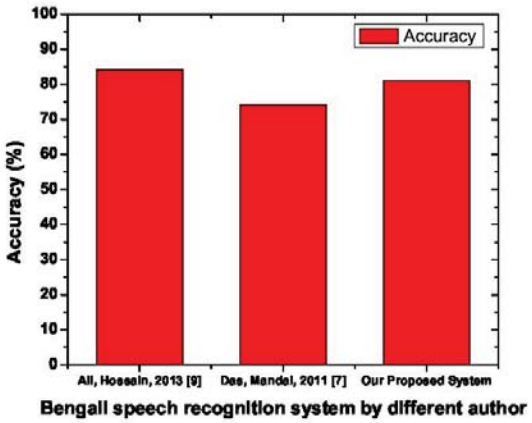

Figure 4. Comparison of accuracy with other recognition system

From figure 4, we can see that our proposed system had better accuracy than paper [5] but had fewer accuracy than paper [7]. But in paper [7], they had recognized only 100 Bengali words.

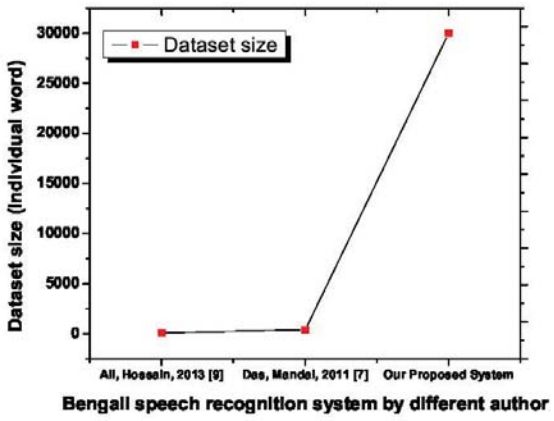

Figure 5. Comparison of dataset with other recognition system

From figure 5, we can see that paper [5] and paper [7] used very few datasets but our proposed system had a huge dataset.

VI. Conclusion

This paper implemented speech recognition system in Bengali language by using two neural networks, namely, convolution neural network and recurrent neural network. The CNN model could convert isolated speech signals into texts and had an accuracy of 86.058%. It indicated that it could provide better predicted output when we used a larger dropout value. We could train the RNN model with 30,000 Bengali words and tried to build an efficient model which converted the speech signals into text transcriptions.

References

[1] The Past, Present, and Future of Speech RecognitionTechnology,"https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf", accessed on 02 November 2018.

[2] Voice Recognition Software An Introduction,"http://www.bbc.co.uk/accessibility/guides/factsheets/factsheet VR intro.pdf", accessed on 02 November 2018.

[3] M. S. Islam, "Research on Bangla language processing in Bangladesh: progress and challenges," 8th International Language & Development Conference, pp. 527-533, 23-25 June 2009, Dhaka, Bangladesh.

[4] R. Gordon, "Ethnologue: Languages ofthe World," 15th Ed., SIL International, Texas, 2005.

[5] Das, S. Mandal and P. Mitra, "Bengali speech corpus for continuous auutomatic speech recognition system," International Conference on Speech Database and Assessments (Oriental COCOSDA), pp. 51-55, 2011, Hsinchu.

[6] M. A. Hasnat, J. Mowla, M. Khan, Isolated and Continuous Bangla Speech Recognition: Implementation, Performance and application perspective, 2007.

[7] M. A. Ali, M. Hossain, M. N. Bhuiyan, Automatic speech recognition technique for bangla words, International Journal of advanced science and technology vol. 50, January 2013.

[8] P. Banerjee, G. Garg, P. Mitra and A. Basu, "Application of triphone clustering in acoustic modeling for continuous speech recognition in Bengali," 19th International Conference on Pattern Recognition,pp. 1-4, 2008, Tampa, FL.

[9] M. A. Hossain, M. M. Rahman, U. K. Prodhan, M. F. Khan, "Implementation of back-propagation neural network for isolated Bengali speech recognition," International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.4, July 2013.

[10] Hannun, C. Case, J. Casper et al. "Deep Speech: Scaling up end-to-end speech recognition," 19 December 2014.