Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

Implementation of Recurrent Neural Network with Language Model for Automatic Articulation Identification System in Bangla

Masiath Mubassira
Department of Computer Science & Engineering, East West University, Dhaka, Bangladesh
Email: masiathmubassira@gmail.com
Amit Kumar Das
Department of Computer Science & Engineering, East West University, Dhaka, Bangladesh
Email: amit.csedu@gmail.com

ABSTRACT

To nudge the state of the art of human-machine interacting applications, research in speech recognition systems has progressively been examining speech-to-text synthesis, but implementation has been done to minimal languages. Although the Bengali language has not been much of an object of interest, we present the automatic speech recognition (ASR) system solely based on this particular language since around 16% of the world's population speak Bengali. It has been a demanding task to implement Bengali ASR because it consists of diacritic characters. We conduct a series of preprocessing and feature selection methods along with a convolutional neural net model in consideration of an automatic verbal communication recognition system. Furthermore, the researchers compared this method to a recurrent neural network that is based on an LSTM network and a vast data file of Google Inc. Investigation of these two models indicates such as the recurrent neural net outperforms the convolutional neural net: the former benefits from combining connectionist temporal classification (CTC) and language model (LM). A quantitative analysis of the output shows that the word error rate and validation loss can be affected by variation in dropout values. It also shows that the parameters are also affected by clean and augmented data.

Keywords - **Convolutional Neural Network, CTC, Word Error Rate, Edit Distance, Augmented Data, Test Loss, Validation Loss, Clean Data, Graphical User Interface**

Date of Submission: Jun 16, 2021 Date of Acceptance: Jun 27, 2021

I. INTRODUCTION

Speech recognition has been one of the enrapturing fields in the study of the interaction between humans and computers. The verbal communication recognition system has been taken up to organize auditory signals that a machine can decipher into text transcriptions. It has penetrated people's consciousness because humans are fascinated by machines that can grasp and sustain them. From an anthropological standpoint, people can utter 150 words in one minute, compared to humans who can type only on average of 40 words in 1 minute. It is predicted that almost half of all searches will be articulation searches by the end of this year [1]. Despite its inhibition, at the starting of the 21st millennium, there is immense voice identification in browsers, the medical sector, service provision, computerized identification, communication in service providers, and mobile phones. Some companies are occasionally bringing new advancements which may create a significant impact in this field. These technologies are of immense help, especially for people with specific disabilities. It has been beneficial for people who are moderate at typing, for those with spelling problems, and also for those who have dyslexia. [2]

Such research on ASR frameworks has been created for various articulations mainly spoken in China, UK, and Germany due to the massive popularity created by such articulations; however, it is still in the prior stage for Bengali language.

paper cannot be confused with a reference [4] or an equation (3) designation. Bengali is the native language of around 245 million people, indicating that around 16% of the world's population speak in Bengali [23] [24] [25]. [3] Bengali is classified as the 7th most used language based on the frequency of Bengali speakers. [4] Thus, there is a compelling requirement for the ASR frameworks to interface with the Bengali language. So this research concentrated mainly on creating a robust system that includes the synthesis of Bengali articulations to sentences.

The Gaussian mixture model (GMM) is considered vital among the most dominant systems which incorporate voice identification system. The Hidden Markov model (HMM) can successfully represent the time-related functioning of acoustic waves by incorporating the states' series. However, deep neural network (DNN) modeling has outperformed the GMM-HMM model only with the new millennium. Due to the tremendous volume of data availability and improved computing capability, DNN has started to evolve as an advanced model. [5]

Convolutional neural network (CNN) has been among the neural nets that can deal with information that has

Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

dimensional structures. CNN can work effectively to identify the image. Moreover, it also showed promising outcomes with articulation identification [26] [27] [28] [29]. Hence, CNN is one of the models which is implemented in this paper.

A considerable amount of systems are non-linear, and their functioning is affected by their current state. Among all the artificial neural networks, the recurrent neural network gives the best identification problems regarding such kinds of systems. The recurrent neural network can treat and store information such as a space-time relationship; thus, it can save the activation values of the neurons for the earlier instance of time. [6] Speech recognition is a system consisting of space-time patterns; therefore, it is rational to implement a method like the recurrent neural net for the Bengali articulation identification system.

This paper's foundational target is implementing the interface between Bengali spoken words and a gadget or a PC via neural networks. The convolutional neural network (CNN) model has been trained with secluded words by utilizing a tool that acquires words in time frames. The paper highlights the CTC and LM techniques that are used to train the recurrent neural networks (RNN) – based systems for Bengali language corpus: a structured aggregation of speech audio files. The systems proposed in this paper use feature extractions (FE); FE is one of the massive progress in ASR frameworks which can identify linguistic contents and discard other information such as emotion, background noise, etc. To complete this undertaking, feature extraction method such as MFCC has been implemented for finding the feature vectors from the acoustic signals.

Our main contribution is, therefore, the development of a Bengali speech recognition system with all the following properties: the proposed model did not need to point out phonemes and to preprocess required to a minimal extent; it supports open-sourced system; it can work offline, and can merge with another system.

The research has been structured in these ways: Section II discusses a few relevant works in Bengali speech recognition, and Section III discusses the experimental setup of the proposed systems. Section IV provides the proposed models, and Section V discusses regularization. Section VI demonstrates the experimental results and discussions, and Section VII presents graphical user interfaces. At last, the conclusion is presented in Section VIII.

II. BACKGROUND

Here a portion of the related works in Bengali speech recognition system was given. Das et al. [7] took a shot for phoneme based system and here 47 phonemes had been picked for the Bengali speech by considering the articulatory system. Forming a speech corpus is important for making and creating any speech recognition system. The automatic speech recognition system had three steps, namely feature extraction, acoustic, and language models. The speech was divided into frames because speech is a continuous signal. MFCC had been utilized to draw out the characteristics from those frames. CMU SPHINX toolkit and HTK were implemented for the auditory system. Monophone system implemented HTK to recognize phoneme and triphone model implemented CMU Sphinx to recognize words. Tri-gram language model had been used for a continuous speech recognition system. Although the correctness proportion declined with information related to the senior citizens, the model worked better with data files related to the young population.

An isolated word speech recognition model is the recognizer built with the concatenation of some distinct words. Mainly, the continuous speech recognition system resembles real speech. Hasnat et al. [8] recorded the speech, and eliminated the noise by using an adaptive filter. The endpoint of a signal was detected by using the endpoint detection algorithm. MFCC was implemented to draw out the characteristics out of the speech waves. After that, two HMM models centered on words and phonemes were utilized to train the isolated-word and continuous verbal identification systems. The secluded word recognizing system could have only identified a word that was present in the vocabulary. It also had the limitation that the speech had to be uttered by the same speaker, and also the speaker's mood had to be similar.

Ali et al. [9] implemented four speech recognition systems and drew out the correlation between every model's elapsed time and recognition rate. Initially, the speech was recorded with a microphone, and then the speech was divided into frames. Background noise was eliminated by using the buffering technique. Pre-emphasis filter was used to eliminate the DC component from the signal. Hamming window and Fourier transform were applied, and finally, feature extraction was used for further analysis. The first system implemented Dynamic Time Warping and MFCC. The second system implemented Dynamic Time Warping and Linear Predictive Coding. The third system implemented the posterior probability function, Gaussian Mixture, and MFCC, which showed the highest accuracy, 84%. The fourth system implemented Dynamic Time Warping, Linear Predictive Coding, and MFCC.

The acoustic model's performance is crucial in evaluating the performance of any continuous speech recognition system. Banarjee et al. [10] drew comparisons between Bengali monophone and triphone-based acoustic models. Since there was a lack of triphone-based acoustic models, triphone clusters had been generated using decision tree techniques. Triphone dependent acoustic system required way more training data than monophone dependent acoustic systems. So it would not be possible that every triphone would be present in the training data. Thus triphones were grouped based on the similarities of phonemes. The outcomes indicate that triphone system accomplished more compared to the monophone centered systems.

Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

Muhammad et al. [11] introduced automatic speech recognition for Bengali digits. The corpus was gathered from the natives in Bangladesh. The sampling frequency of 16-bit sample resolution, first and second-order derivatives of 13 MFCC parameters, and Hamming window consisting of a step size of 10 milliseconds were the parameters used for this paper. The automatic Bengali speech recognition model had been created by implementing the hidden Markov model toolkit (HTK). They selected 50 males and 50 females, a sum of 100 orators to build speech vocabulary. The orators' ages were between 16 and 60 years. This paper divided the speech corpus, such as 37 males and 37 females were training sets, and the rest were testing sets. The highest correct rate was for digit '২' (2), and the lowest correct rate was for '৮' (8). The most two confused pairs were '৬' (6) and '৯' (9); and '৭' (7) and '৮' (8).

Hossain et al. [12] recorded all the Bengali digits uttered by 10 speakers. The characteristics of the speech from 5 speakers were drawn out by implementing MFCC. These characteristics had been utilized to instruct the system, which consisted of Backpropagation neural net. The numbers spoken by the other 5 orators had been utilized to check the framework. The system acquired an identification frequency of around 96% in the case of familiar orators and around 93% in the case of unfamiliar orators.

Microsoft Corporation built Speech Application Program Interface (SAPI), which incorporated articulation relevant systems for Windows Operating System. It included characteristics that are available for 8 mother tongues spoken mainly in USA, UK, France, Spain, Germany, Japan, and China. So Sultana et al. [13] implemented the synthesis of articulation to words based on SAPI for Bangla speech. The research matched words from continuous Bengali speech with speech corpus of SAPI, and if both were matched, SAPI returned the Bengali words in English character. These words were then used to return the Bengali words from the database in Bengali characters. On average, the recognition rate for this paper was approximately 78%.

Bhowmik et al. [14] proposed the traditional approach; the phoneme-based model was used for the recognition system. Stacked Denoising Autoencoder was utilized for training the deep neural network, which consisted of 3 hidden layers; each layer had been utilized with 200 numbers hidden units. The updated probability of the phonemes was acquired in the output layer. Data was collected from many sources such as Radio, Television, and conversation in the laboratory.

Manjunath et al. [15] explored the integration of MFCC and speech characteristics for developing tandem mobile identification models for TIMIT and Bangla articulation corpus. HMM, and DNN was used for developing the mobile identification models where DNN had outperformed HMM in each case. The phone-posterior-based and speech characteristics centered tandem mobile identification models had revealed largest mobile identification precision for TIMIT and Bengali data files.

Phadikar et al. [16] developed a method to identify isolated Bengali alphabets from acoustic signals. MFCC is a better representative of the human auditory system, but performance deteriorates with increased noise; however, it may be decreased by using Discrete Wavelet Transform (DWT). The database was developed by ten speakers who uttered 43 alphabets, and each alphabet had been uttered 20 times. The features extracted were used to train many classifiers from the Weka tools. The grouping accuracy was studied by implementing ten-fold cross-validation method.

Tripathi et al. [17] developed the speech mode classification model, which contained three forms of speech. The characteristics of the voice were drawn out by implementing MFCC; the excitation features were captured through MPDSS and RMFCC. Classification models such as k-nearest neighbor (KNN), support vector machines (SVM), naïve Bayes, as well as artificial neural network (ANN) were used for analyzing the speech mode classification model.

Sumon et al. [18] implemented CNN based system on a small dataset. Three approaches were implemented; the first approach used MFCC to extract features for training the CNN model, the second approach trained the model by using raw audio signals, and the third approach extracted features by using transfer learning.

Sabab et al. [19] implemented 1-dimensional CNN and MFCC which showed better accuracy than the MFCC and LSTM model.

III. EXPERIMENTAL SETUP

A. Speech Corpus

1) Speech Corpus for CNN System

The dataset is comprised of articulation data of the extension wav incorporating recitals in regard to some words 'bangla', 'gaan', 'shundor', and numerals from 1 to 3 in the Bengali language. Each word is uttered 35 times on an average by 2 speakers; however, it is mostly trained with a single speaker.

2) Speech Corpus for RNN System

The Bengali speech data is collected from Google Inc. The speech data are spoken by around 510 speakers. Although the corpus consists of around two-hundred thousand articulation clips, at most, around 35000 articulation clips have been implemented in regard to this paper. The total corpus is distributed for training, testing, and validation; 80 percent utilized in the form of training values, 90 percent utilized in the form of validation values, 10 percent utilized in the form of test data. The complete time length in regard to around 35000 data files had been around 1980 minutes. The audio data files were transformed into documents with the extension wav. The sampling frequency is around 17000 Hz, along only 1 channel is implemented. The text clip is 16-bit Pulse Code Modulation (PCM) encrypted transcripts. The datasets

Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

used are OpenSLR37 and openSLR53. The model is trained by using Google Collab.

The dataset contains diacritic characters; a diacritic is a sign used with an alphabet that produces a different pronunciation compared to that of the alphabet itself. A vowel can follow any consonant in Bengali, that is, a diacritic. For example, the consonant 'অ' can be followed by the diacritic 'া'. Thus the Bengali script becomes 'আ' [20]. Therefore it is more complicated to convert Bengali audio clips into text transcriptions because it consists of diacritic characters. So building a Bengali speech recognition system is more difficult compared to some other languages [30] [31] [32] [33].

B. Input Features

Feature extraction is used to calculate some parameters by processing the audio waveform. A most popular spectral analysis technique is the Mel Frequency Cepstral Coefficient (MFCC) which was utilized in the form of input in the neural network models. MFCC extract speech features which are analogical to the way humans perceive and filter speech. The following procedures have determined MFCC: initially, the articulation wave is split into a number of small samples. The power spectrum estimation is determined for every frame, and the mel filter bank is implemented to the power spectrum. The logarithm is calculated of all filter bank energies. The Discrete Cosine Transform (DCT) is extracted from the log filter bank energies. 2 to 13 DTC coefficients are used, and the rest are eliminated.

C. Softwares

The Bengali ASR system has been built by implementing Python programming language. The Python deep learning library used for both neural networks is Keras. The backend engine used for Keras is Tensorflow. For making the speech corpus for CNN based model, the Udacity tool has been implemented to acquire the speech frames comprised of 1 second each.

D. Hardware System

For CNN system, an ASUS Precision X556U workstation has been used for the experimental procedure. The machine used has a single core intel xeon processor with a 2.3 GHz CPU clock speed, cache memory of 45 MB and RAM of 12 GB, and 1xTesla K80 Graphical Processing Unit (GPU).

IV. PROPOSED MODEL

We implemented two models based on deep neural networks because these models performed well and gave great accuracies for speech recognition systems compared to probabilistic models.

E. CNN Based Architecture System

The architecture of CNN consists of three layers: fully connected layer, downsampling layer, and convolutional layer. The convolutional layer incorporates a collection of learnable neurons. During a forward pass, the activation map is computed by multiplying the receptive field and weights of the neurons. The pooling layer used in this paper is maximum pooling. It takes the highest amount in a vector neighborhood. Thus the pooling layers take the best matching inputs. The fully connected layer connects each neuron of all the intermediate layers. The activation map is passed to the fully connected layer for classification.

The activation functions used for the convolutional layer and the output are rectified linear unit (ReLU) and softmax, respectively. ReLU, a non-linear activation function, is not affected by the vanishing gradient problem. Softmax is used to classify the output at the fully connected layer. The number in this activation function represents the probability for a particular class.

The data use forward propagation for proceeding in the forward direction. Prior to convolving the speech files, the biases as well as weights have been assigned for nodes along with edges. Multiple convolutional and pooling layers can be used to process the data. The activation function normalizes the features in different layers. Backpropagation is used to figure out the extent to which the neural network needs adjustment. Gradient descent is calculated, and therefore, the weights and biases are adjusted.

F. RNN Based Architecture System

1) Formulation of RNN

The RNN was trained to detect speech spectrograms as inputs and was converted into Bengali text transcriptions. There were five layers in the neural network. Three completely connected layers were given the input, the following layer is a directional RNN layer, and the last layer is a completely connected layer. The activation function used for the fully connected hidden layer was rectified linear unit (ReLU). The activation function appropriate to the RNN layer was tanh, and Long-Short-Term-Memory (LSTM) cells were also had been used. Let one utterance be considered as u along with the output be considered as v in regard to a training group

$$Y=\{(u(1),v(1)),(u(2),v(2)),\} \quad (1)$$

Each articulation, u(j), contained a duration of extent C(j) in which every duration scale contained the vector consisting of articulation characteristics, uc(j) c=1, 2, …., C(j). The objective of the recurrent neural network was to transform the input series to a series of probabilities corresponding to each character for the audio transcripts, v with vc=P(xc|u) in which xc is the member of {a,b,c,d,….,z, blank, space,apostrophe}.

For an input a, the layer k had many hidden units which can be represented by m(k), and the input can be denoted by m(o). The first three layers of the network were not recurrent. In the first layer, at time t, the outcome relied upon the spectrograph sample u(t) as well as relied upon the former along with latter samples of u(t). The remainder of the non- recurrent layers operated on non-dependent data for every time slice. Hence, for each time t, the former 3 layers were calculated by:

$$mt(n) = h(E(n)mt(n-1)+b(n)) \quad (2)$$

Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

in which the rectified-linear (ReLu) was represented by h(y) = minimum{maximum{0,y},20} and b(n), E(n) were the bias parameters and weight matrix for layer n respectively.

Bidirectional RNN layer was the immediate layer. The layer contained 2 groups consisting of concealed portions: a group that consisted of forward propagation net, m (fp), along with the group that consisted of backward propagation net m (bp):

$$mt(fp)= h(E(4)mt(3) + Er(fp)\ mt\text{-}1(fp) + b(4)\ ) \quad (3)$$

$$mt(bp) = h(E(4)mt(3) + Er(bp)\ mt\text{+}1(bp) + b(4)) \quad (4)$$

It should be pointed out that m(f) had to be calculated consecutively for the j'th utterance in between time= 1 and time = T(j). However, these units m(b) were calculated between time=T(j) and time=1 consecutively by following the opposite sequence. The 5th layer got intake via the forward unit as well as from the backward unit. The output contained the softmax function that produced output in the form of character probability with regard to every character of the alphabet.

2) Enhancing Precision by Implementing CTC Along with Language Model

After the research calculated an estimation with regard to each character in the speech transcript, the CTC loss is computed for evaluating the error in the prediction of the transcription. This paper updated the neural network's model parameters and tried to minimize the loss function by using back-propagation recurrent network.

For lots of cases, RNN can perfectly predict the character sequence of an audio file. The problem arises when the predicted words never or occasionally occurred in the dataset. One solution is to collect utterance for as many words as possible; however, this solution is impractical. So it is more efficient to use a language model. Considering the outcome of the neural network, the model tried to figure out the most probable series of characters in relation to the RNN output and also the language model in which the language model gave an output of the sequence of characters in terms of words. Explicitly, this paper found a series q that would have maximized the integrated objective:

$$S(q) = \lg (P(q|a)) + \beta \lg (Plm(q)) + \alpha \text{ word count}(q) \quad (5)$$

in which β along with α are adjustable weights that had been assigned using cross-validation. This checked the balance of the language model with the recurrent neural net. The notation Plm indicated the likelihood of speech transcript. [21]

V. REGULARIZATION

Neural networks generally have many trainable parameters, which may cause overfitting; i.e., these networks perform well with training data, but performance deteriorates with unseen data. Regularization focuses on improved generalization by minimizing overfitting in many ways. The most efficient method for avoiding overfitting is dropout [22], where hidden nodes are dropped randomly. Thus dropout has been used for the recurrent neural network model.

VI. RESULTS AND DISCUSSION

G. Result and analysis drew on CNN model

The paper trained the model by using 30 samples of each word. No background noise was added to the samples, and the model was trained with raw data. Then for testing this recognition system, a vocabulary consisting of a different sample was taken. Then the model was used for recognizing the Bengali spoken words.

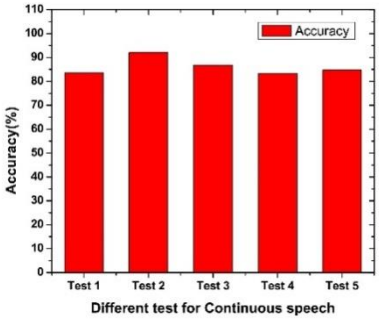

Fig. 1.Accuracy for the 5 tests conducted for speech recognition system (speaker dependent)

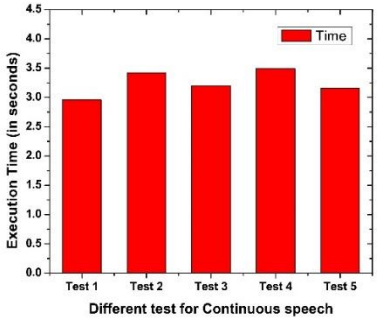

Fig. 2.Required time to run the model in the 5 tests (speaker-dependent)

Figure 1 indicates that we have run five different test cases, and we have found an average of 86.058% accuracy. The required time to recognize words was 3.246 seconds on average.

In figure 2, we have tested our model for five different data sets. Then we have computed the required time to recognize the speech for each test. We have found that the average time to recognize the speech is 3.24 seconds, where the minimum required time is 2.96 seconds, and the maximum required time is 3.49 seconds. The standard deviation is 0.19.

Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

H. Result and Analysis Based Upon RNN System

This execution is assessed with respect to test loss, validation loss, dropout values, edit distance, word error rate, clean data, augmented data, variation in hidden cells, language model weight, and word count. For a few of the experiments, comparisons are made concerning dropout. These hyperparameters are altered as well as the updated results are noted. Generally, the concealed cells used for the recurrent neural net are around 2050. The step size is around 0.0002. In every analysis, a summation of around 55 epochs is implemented. The accuracy for RNN system is 85%.

1) Loss Comparison among Various Dropout Values

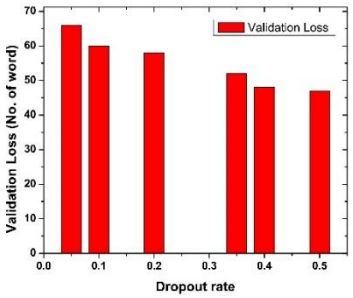

Fig. 3. Validation loss change with dropout

The dropout values are changed with respect to validation loss, as shown in figure 3. For figure 3, in horizontal axis dropout values are considered, and in vertical axis validation, the loss is considered. It can be demonstrated that greater values of dropout accomplish lower validation loss. Figure 3 also indicates that as we use lesser parameters, the model becomes more accurate.

2) Comparison of Loss between Clean Data and Augmented Data for Variation in Concealed Cells

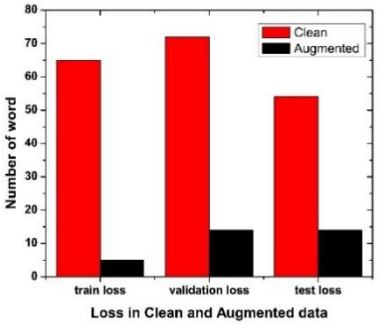

Fig. 4. Comparison of loss between clean data and augmented data when the value of concealed cells is 1024

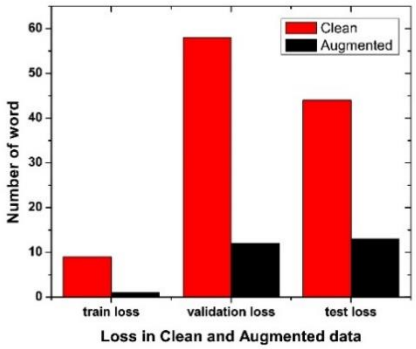

Fig. 5. Comparison of loss between clean data and augmented data when the value of concealed cells is 2048

The concealed cells intended for figure 4 and 5 are 1024 and 2048 respectively. For figures 4 and 5, if we compare clean data with augmented data, test loss, train loss, and deviation loss are higher. This shows that if data is augmented, if the speed, frequency, and rest of the properties of the speech sample are altered, thus the system can come up with a finer-estimated output of the transcription. However, train loss for clean data for 2048 hidden cells is lesser than that of the dataset containing hidden cells of 1024.

3) Comparison of Test Edit Distance Change and Test Word Error Rate with Variation in Dropout Values

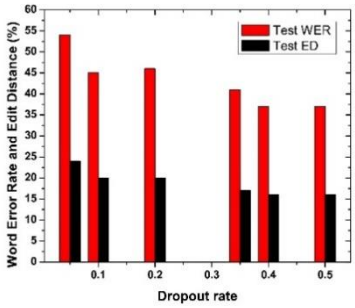

Fig. 6. Comparison of test edit distance(ED) and test word error rate (WER) change with changes to dropout values

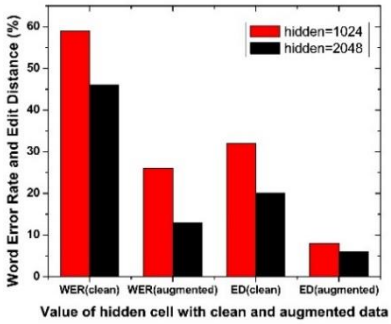

Fig. 7. Edit distance change and word error rate on the test set

Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

In figure 6, it is evident that the edit distance rate and word error rate and decrease with an increase in the dropout rate. This shows that the predicted output is more accurate if the dropout value is decreased.

From figure 7, we can see that for both hidden cells of RNN, edit distance rate and word error rate for clean data are more than that of augmented data.

4) Comparisons of Dataset of Various Works

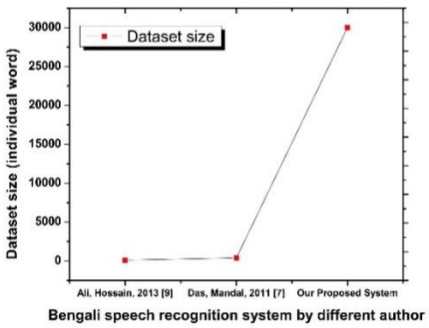

Fig. 8. Comparison of the data samples of other models

In figure 8, we can observe that Ali 2013 and Das 2011 utilized minimal datasets, but this paper's proposed system has a tremendous volume of the dataset.

5) Test Word Error Rate Comparison for Variation in Language Model Weight and Word Count

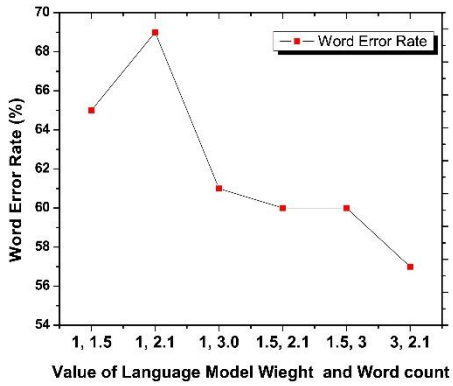

Fig. 9. Test WER comparison on different language model weights

From figure 9, we can observe that there is not any strong correlation between WER and language model weights. If the language model weight remains constant and word count increases, the WER increases, then it decreases and then gets constant to a certain value.

6) Comparison of Test Edit Distance and Word Error Rate for Various Datasets

TABLE I. COMPARISON OF TEST WER AND EDIT DISTANCE FOR DIFFERENT DATASETS

Dataset	***WER***	***Edit Distance***
OpenSLR 53 (20%)	0.6	0.28
OpenSLR 37 (Full)	0.37	0.16
OpenSLR 37 (Full Augmented)	0.19	0.06

From table 1, we can observe that WER and edit distance decrease if we use augmented data for the dataset OpenSLR37. OpenSLR53 has more WER and edit distance than that of OpenSLR37.

VII. GRAPHICAL USER INTERFACE

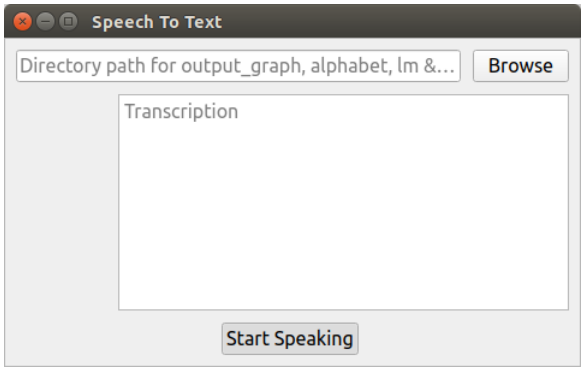

Fig. 10. Graphical user interface for speech to text synthesis

Figure 10 presents a graphical user interface that is used for taking speech signals as input, and our RNN model has been used to find text transcriptions. Some speech-to-text synthesis is given as examples below.

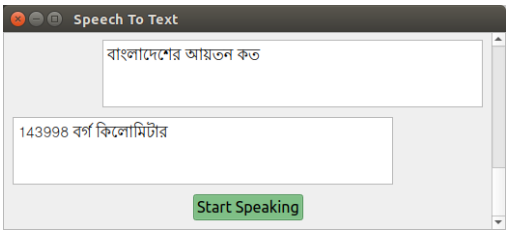

Fig. 11. Speech to text synthesis showing Bangladesh's area

From figure 11, we can see that the speech is processed and converted into text transcription in Bengali, which means 'What is the area of Bangladesh?' The user interface responds and shows the text transcription in Bengali, which means '143998 square kilometers.'

Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

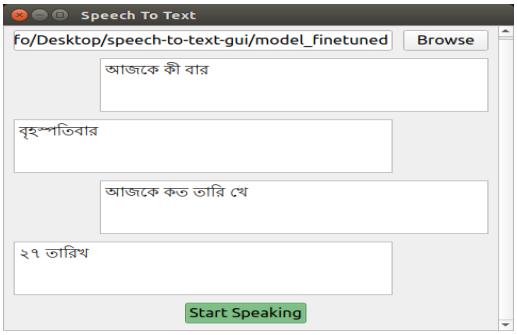

Fig. 12. Speech to text synthesis showing date and day

From figure 12, we can observe that the speech is processed and converted into text transcription in Bengali, which means 'What day is today?' The user interface responds and shows the text transcription in Bengali, which means 'Thursday.' It again shows another text transcription which means 'What date is it today?' It gives the result which indicates '27th.'

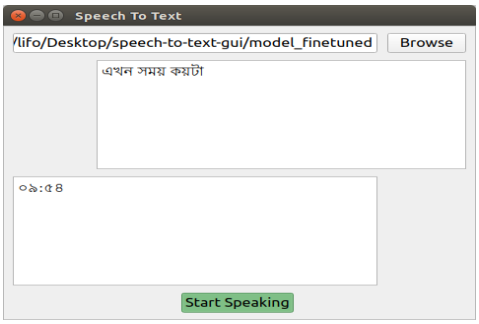

Fig. 13. Speech to text synthesis showing time.

From figure 13, we can demonstrate that the speech is processed and converted into text transcription in Bengali, which means 'What time is it?' The user interface responds and shows the text transcription in Bengali, which means '09:54.'

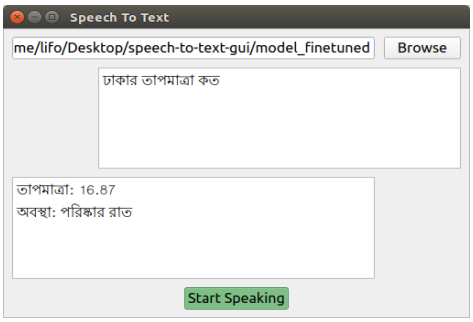

Fig. 14. Speech to text synthesis showing temperature in Dhaka

From figure 14, we can see that the speech is processed and converted into text transcription in Bengali, which means 'What is the temperature in Dhaka?' The user interface responds and shows the text transcription in Bengali, which means 'Temperature: 16.87 Condition: Clear night.'

VIII. CONCLUSION

Lately, the neural network has accomplished itself as an improved method for a speech recognition system. Since there is an imperative requirement for ASR framework to be created in Bengali language, such a vital exertion has been acquired for Bengali spoken sentences in this work. The spontaneous speech system is trained by using CNN and RNN models. In the RNN based ASR system, CTC and language model are utilized to extricate an efficient output. The experimental analysis demonstrates that the ASR system performs better with dropout and augmented data. In the future, the neural network will be trained with more data and more augmentation; it will be integrated with other applications, and other neural network architectures will also be implemented. Also, background noise will be added to the dataset to observe the effect of noise in the study.

REFERENCES

[1] The Past, Present, and Future of Speech Recognition Technology,"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", in proc. of 8th International Language & Development Conference, pp. 527-533, 23-25 June 2009, Dhaka, Bangladesh.

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

[5] T. Aditi, V. Karun, "Speech recognition of Punjabi numerals using convolutional neural networks", Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing, vol 759, Springer, Singapore.

[6] Mon A.N., Pa W.P., Thu Y. K. (2018) Exploring the Effect of Tones for Myanmar Language Speech Recognition Using Convolutional Neural Network (CNN). In: Hasida K., Pa W. (eds) Computational Linguistics. PACLING 2017. Communications in Computer and Information Science, vol 781. Springer, Signapore.

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

[8] Md. A. Hasnat, J. Mowla, M. Khan, "Isolated and continuous Bengali speech recognition: implementation, performance and application perspective," 2007.

[9] Md. A. Ali, M. Hossain, M. N. Bhuiyan, "Automatic speech recognition technique for bangla words,"

Volume: 12 Issue: 06 Pages: 4800-4808(2021) ISSN: 0975-0290

International Journal of advanced science and technology vol. 50, January, 2013.

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

[11] G. Muhammad, Y. A. Alotaibi, M. N. Huda, "Automatic speech recognition for Bangla digits," 2009 12th International Conference on Computers and Information Technology, Dhaka, 2009, pp. 379-383.

[12] Md. A. Hossain, Md. M. Rahman, U. K. Prodhan, Md. 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.

[13] Sultana, Shaheena, Akhand, M. A. H., Das, Prodip, Rahman, M. M.. (2012). Bangla Speech-to-Text conversion using SAPI. 385-390. 10.1109/ICCCE.2012.6271216.

[14] Bhowmik, Tanmay, Choudhury, Amitava, Mandal, Das. (2018). Deep Neural Network Based Recognition and Classification of Bengali Phonemes: A Case Study of Bengali Unconstrained Speech.

[15] Manjunath, K.E., S. Rao, K. Circuits Syst Signal Process (2018) 37: 704.

[16] Phadikar S., Das P., Bhakta I., Roy A., Midya S., Majumder K. (2017) Bengali Phonetics Identification Using Wavelet Based Signal Feature. In: Mandal J., Dutta P., Mukhopadhyay S. (eds) Computational Intelligence, Communications, and Business Analytics. CICBA 2017. Communications in Computer and Information Science, vol 775. Springer, Singapore.

[17] Tripathi, K., Rao, K.S. Int J Speech Technol (2018) 21: 489

[18] S. Ahmed Sumon, J. Chowdhury, S. Debnath, N. Mohammed and S. Momen, "Bangla Short Speech Commands Recognition Using Convolutional Neural Networks," 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018, pp. 1-6, doi: 10.1109/ICBSLP.2018.8554395.

[19] Sabab M.N., Chowdhury M.A.R., Nirjhor S.M.M.I., Uddin J. (2020) Bangla Speech Recognition Using 1D-CNN and LSTM with Different Dimension Reduction Techniques. In: Miraz M.H., Excell P.S., Ware A., Soomro S., Ali M. (eds) Emerging Technologies in Computing. iCETiC 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 332. Springer, Cham. https://doi.org/10.1007/978-3-030-60036-5_11

[20] Vowel Diacritics in Bengali," https://en.wikibooks.org/wiki/Bengali/Script/Diacritics", accessed on 02 November 2018.

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

[22] Schluter R. et al. (2016) Automatic Speech Recognition Based on Neural Networks. In: Ronzhin A., Potapova R., Nemeth G. (eds) Speech and Computer Science, vol 9811. Springer, Cham.

[23] O. F. Rakib, S. Akter, M. A. Khan, A. K. Das and K. M. Habibullah, "Bangla Word Prediction and Sentence Completion Using GRU: An Extended Version of RNN on Ngram Language Model," 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), 24-25 December, Dhaka.

[24] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das and T. Mittra, "A Deep Learning Approach to Detect Abusive Bengali Text," 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019.

[25] M. M. Hossain, M. F. Labib, A. S. Rifat, A. K. Das and M. Mukta, "Auto-correction of English to Bengali Transliteration System using Levenshtein Distance," 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019.

[26] M. D. Drovo, M. Chowdhury, S. I. Uday and A. K. Das, "Named Entity Recognition in Bengali Text Using Merged Hidden Markov Model and Rule Base Approach," 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019.

[27] E. Biswas and A. K. Das, "Symptom-Based Disease Detection System In Bengali Using Convolution Neural Network," 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019.

[28] A. K. Das, A. Ashrafi and M. Ahmmad, "Joint Cognition of Both Human and Machine for Predicting Criminal Punishment in Judicial System," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 36-40.

[29] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter and A. K. Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 360-364.

[30] J. Islam, M. Mubassira, M. R. Islam and A. K. Das, "A Speech Recognition System for Bengali Language using Recurrent Neural Network," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 73-76.

[31] T. F. Mumu, I. J. Munni, and A. K. Das, "Depressed People Detection from Bangla Social Media Status using LSTM and CNN Approach", J. eng. adv., vol. 2, no. 01, pp. 41-47, Mar. 2021.

[32] M. T. Hossain, M. W. Hasan and A. K. Das, "Bangla Handwritten Word Recognition System Using Convolutional Neural Network," 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), 2021, pp. 1-8.