PERFORMANCE ANALYSIS OF DIFFERENT ACOUSTIC FEATURES BASED ON LSTM FOR BANGLA SPEECH RECOGNITION

Nahyan Al Mahmud

Department of Electrical and Electronic Engineering,
Ahsanullah University of Science and Technology, Dhaka, Bangladesh

ABSTRACT

In this work a new Bangla speech corpus along with proper transcriptions has been developed; also various acoustic feature extraction methods have been investigated using Long Short-Term Memory (LSTM) neural network to find their effective integration into a state-of-the-art Bangla speech recognition system. The acoustic features are usually a sequence of representative vectors that are extracted from speech signals and the classes are either words or sub word units such as phonemes. The most commonly used feature extraction method, known as linear predictive coding (LPC), has been used first in this work. Then the other two popular methods, namely, the Mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) have also been applied. These methods are based on the models of the human auditory system. A detailed review of the implementation of these methods have been described first. Then the steps of the implementation have been elaborated for the development of an automatic speech recognition system (ASR) for Bangla speech.

KEYWORDS

Mel frequency cepstral coefficients, linear predictive coding, perceptual linear prediction, sentence correct rates, LSTM.

1. INTRODUCTION

Speech is a versatile mean of communication. It is the most natural, flexible, efficient and convenient means of conveying linguistic, speaker identification and environmental information. Even though such information is encoded in a complex form, humans can relatively decode most of it. This human ability has inspired researchers to develop systems that would emulate such ability. From phoneticians to engineers, researchers have been working on several fronts to decode most of the information from the speech signal. Some of these fronts include tasks like speech recognition, speaker recognition, voice analysis (for medical purposes), speech synthesis, speech enhancement, speech compression, speaker logging, language detection, translating speech, and understanding speech.

There have been many literatures in automatic speech recognition (ASR) systems for almost all the major languages in the world. Unfortunately, only a very few work have been carried out in automatic speech recognition (ASR) for Bangla, which is one of the largely spoken languages in the world. A major difficulty in conducting research in Bangla ASR is the lack of proper speech corpus. Some efforts have been made in the past to develop Bangla speech corpus to build a Bangla text to speech converter [1].

Some developments on Bangla speech processing can be found in references [2]-[8]. Continuous Bangla speech recognition systems are developed in [2]-[5]. Bangla vowel characterization is accomplished in [6], Bangla speech recognition on a small dataset using hidden Markov models (HMMs) is described in reference [6], whereas recognition of Bangla phonemes by artificial neural network (ANN) is reported in [7]-[8], while [5] presents a brief overview of Bangla speech synthesis and recognition. However, most of these work mainly concentrated on simple recognition tasks on a very small database and did not consider the colloquial effect in different parts of the country.

The main objectives of this research are:

(a) To develop a database/corpus for Bangla speech recognition system.

(b) To investigate the effectiveness of different acoustic features for Bangla speech recognition using LSTM neural network and evaluate their performances based on the spoken Bangla words.

The outcomes of the current research may be as follows:

A database for Bangla speech has been developed which can be used for speech analysis. Also, the effectiveness of an LSTM-based speech recognition system with various acoustic features for Bangla speech has been assessed.

2. AUTOMATIC SPEECH RECOGNITION FOR BANGLA SPEECH

2.1. Bangla Speech Corpus

At present, a real problem to carry out experiment on Bangla ASR is the lack of proper Bangla speech corpus. In fact, such a corpus with a sufficiently large database is not available or at least not referenced in any of the existing literature. On the other hand, a medium size speech corpus for Bangla digits has been designed, where '0' [Zero] to '9' [Nine] are recorded in reference [9]. They selected 50 male and 50 female subjects with a total of 100 speakers. A total of 50 utterances were recorded from each speaker in quiet office environment. This digit corpus is medium in size, but it is designed only for digit recognition. Besides, the large scale Bangla speech database designed in [10] is not segmented or labelled for use in supervised learning.

Hundred common sentences from the Bengali newspaper are uttered by 100 different speakers of different regions of Bangladesh. Among them, (50x100) sentences uttered by 50 male speakers are used as the male training corpus. On the other hand, the same (50x100) sentences uttered by 50 female speakers are used as the female training corpus. Moreover, a separate set of 100 sentences uttered by 10 different male and 10 different female speakers are used as male test corpus and female test corpus, respectively. Each of these sentences contains two, three or four monosyllabic and polysyllabic Bangla words.

All of the speakers are Bangladeshi nationals and native speakers of Bangla. The age of the speakers ranges from 20 to 30 years. The speakers have been chosen from a wide geographical area of Bangladesh. Though all of them speak in standard Bangla, they are not free from their regional accent.

2.2. Recording Environment and Specifications

Recording was done in a quiet room located at Ahsanullah University of Science and Technology (AUST), Dhaka, Bangladesh. A 15 feet by 17 feet room having styrofoam suspended ceiling was used as a recording room. The room was surrounded by regular office curtain. A dynamic microphone was used for training process and a smart phone was used for testing in order to reflect the real-world situation. Recording was done in an environment with regular ambient noise e.g. noise from ceiling fan and air conditioner and some street or corridor noise.

An unidirectional microphone was used for this work, which has the following specifications.

Type:	Dynamic
Directivity:	Unidirectional
Sensitivity:	-76 dB at 1 kHz
Impedance:	600 ohms
Frequency Response:	40 – 16 kHz

Audacity 2.4.0 software [11] was used for recording. The speech was sampled at 16 kHz and quantized to 16 bit stereo coding without any compression and no filter was used on the recorded voice. Waveform Audio File (WAV) [12] is a suitable file format for uncompressed recording. The speakers were instructed to speak at normal speed and to leave normal gaps between words. Each sentence was recorded in a separate WAV file and sequential file names were assigned to facilitate tracking of all the files. It is not necessary to synchronize all the recordings for the same sentence as the HTK was used to align each word and LSTM is capable of detecting a silence.

2.3. Development of Bangla ASR

The first step in building a dictionary is to create a sorted list of the required words. However, it is necessary to build a word list from the sample sentences present in the training data. Furthermore, to build a robust acoustic models, it is necessary to train it on a large set of sentences containing many words which should be phonetically balanced.

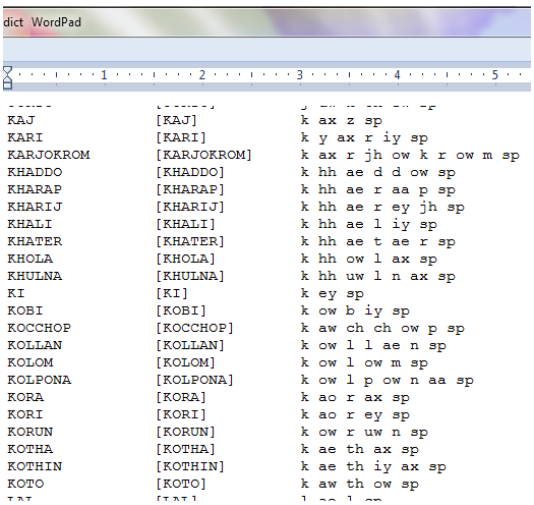

Figure 1. Segment of developed dictionary file

The pronunciation of a word can be given as a series of symbols that correspond to the individual units of sound that make up a word. These are called 'phonemes' or 'phones'. A monophone

refers to a single phone. Thus a monophone model for Bangla can be created from Table 1 and Table 2 along with some help from reference [13]. A silence /sil/ and short pause /sp/ monophone is added to create a silence model. An additional monophone /ax/ is added to indicate a short 'AA', where /aa/ indicates long 'AA' [14].

Table 1. Some Bangla Words With Their Orthographic Transcriptions and IPA

English Pronunciation	IPA	Used Symbol
AAMRA	/a m ɾ a/	/aa m r ax/
AACHORON	/a ʧ ɾ n/	/aa ch ow r aa n/
ABEDON	/a b æ d̪ n/	/ax b ae d aa n/

Table 2 lists some Bangla words with their written forms and the corresponding IPA. From the table, it is shown that the same 'AA' (/a/) has different pronunciation based on succeeding phonemes. These pronunciations are sometimes long or short. For long and short 'AA' we have used two different phonemes /aa/ and /ax/, respectively.

A tri phone model is derived to address the context independency of the mono phone model. A tri phone is simply a group of 3 phones in the form "L-X+R" - where the "L" phone (i.e. the left-hand phone) precedes "X" phone and the "R" phone (i.e. the right-hand phone) follows it. Table II shows an example of the conversion of a mono phone declaration of two words "BANGLADESH" and "SHANTI" to a tri phone declaration.

Table 2. Converting a mono phone declaration to a tri phone declaration

English Pronunciation	Mono phone Declaration	Tri phone Declaration
BANGLADESH	b aa ng l aa d eh sh	b+aa b-aa+ng aa-ng+l ng-l+aa l-aa+d aa-d+eh d-eh+sh eh-sh
SHANTI	sh ax n t ih	sh+ax sh-ax+n ax-n+t n-t+ih

Similarly, we have considered all variations of same phonemes and consequently, found total 51 phonemes excluding beginning and end silence (/sil/) and short pause (/sp/).

2.4. Acoustic Features

Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters. The first ideas leading to LPC started in 1966 [15].

The most widespread acoustic features are mainly based on models of the human auditory system. Some constraining properties of the human hearing such as nonlinear (Bark or Mel) frequency scale, spectral amplitude compression, decreasing sensitivity of hearing at lower frequencies, and large spectral integration are already integrated in the state-of-the-art acoustic features for automatic speech recognition. Another frequently used acoustic features; the Mel Frequency Cepstrum Coefficients (MFCC) feature was first introduced in [15]. The Perceptual Linear Predictive (PLP) feature introduced in [16] is based on ideas similar to the MFCCs. Nevertheless, there are major differences in data flow and in recognition performance as well.

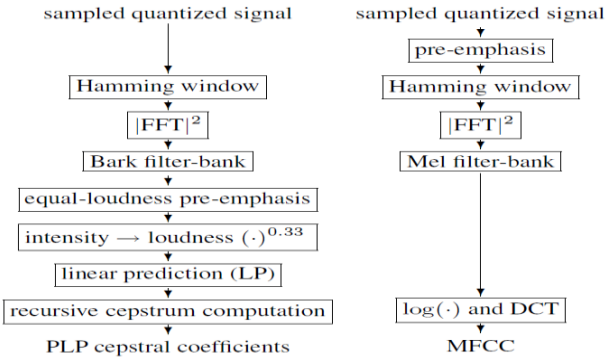

Figure 2. The computation steps of PLP (left) and MFCC (right).

3. EXPERIMENTAL SETUP

3.1. LSTM Neural Network

In deep neural network (DNN) based speech recognition, it is assumed that the sequence of observed speech vectors corresponding to each word is generated by a LSTM model as shown in Figure 4. In the traditional DNN, the amount of residual index that needs to be returned diminishes, when the time is long, which slows down the updating of the network [17]. Long Short-Term Memory (LSTM) solves this by taking advantage of its relatively long-term memory [18].

Long Short-Time Memory (LSTM), a special kind of Recurrent Neural Network (RNN), which can be able of learn long-term dependencies. Remembering information for long periods of time is LSTM's specialization [19]. LSTM models outperform RNN in learning context-free and context sensitive languages [20]. The LSTM speech detection block is illustrated in Figure 3.

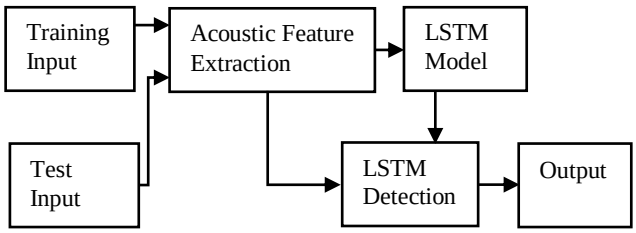

Figure 3. LSTM speech detection block

A deep LSTM - a stack of multiple LSTM layers - combined with a Connection Temporal Classification (CTC) output layer, is applied on the acoustic model [19,20].

3.2. HTK Toolkit

The input is divided into two parts: label and acoustic feature extraction. Though LSTM neural network model is used in this work, htk toolkit is only used for acoustic feature extraction. After obtaining this features namely LPC, MFCC and PLP, these are then trained by LSTM model. The test set is then processed by the same manner. Then the acoustic model decodes the above mentioned features, and finally performs the speech recognition to get the sentence correct rate.

In this paper, the performance of LPC, MFCC and PLP in LSTM neural network structures on Bangla speech recognition will be compared.

4. LSTM STRUCTURE

In a standard structure of LSTM neural networks, there is an input layer, a recurrent LSTM layer and an output layer. In this experiment the Time Delay Neural Network (TDNN) layer is the input layer which is connected to the hidden layer. The recurrent connections are directly from the cell output units to the cell input units, input gates, output gates and forgot gates.

We used the new LSTM structure proposed by Google [19], through which Google has greatly improved the speech recognition ability based on this CTC training. We used this CTC structure to enhance the performance of this setup shown in Figure 4. We experimented with a variety of residual network structures and found that this network has a balanced performance on large data sets, thus increasing the input of each layer and the generalization ability and the following experiments are using this structure of LSTM.

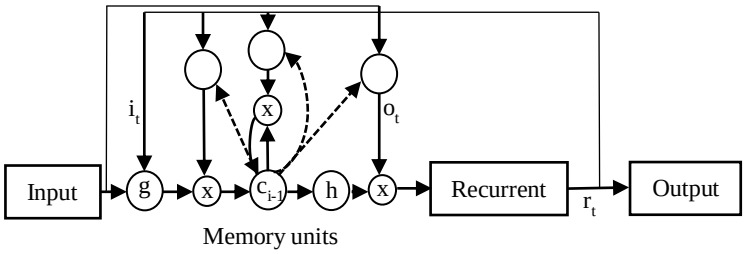

Figure 4. LSTM speech detection model

$$C_t = f_t \odot C_{t-1} + i_t \odot \mathrm{g}(W_{cx} x_t + W_{cr} r_{t-1} + b_c) \ldots (1)$$

$$i_t = \sigma(W_{ix} x_t + W_{ir} x_{t-1} + b_i) \ldots (2)$$

$$o_t = \sigma(W_{ox} x_t + W_{or} r_{t-1} + W_{oc} c_t + b_o) \ldots (3)$$

$$r_t = W_{rm} m_t \ldots (4)$$

It can be seen that equations (1)-(4) are standard formulas for LSTM neural network model.

Continuous speech simply involves connecting models together in sequence. Each model in the sequence corresponds directly to the assumed underlying symbol. These could be either whole words for so-called connected speech recognition or sub-words such as phonemes for continuous speech recognition. The reason for including the non-emitting entry and exit states should now be evident, these states provide the glue needed to join models together.

5. PERFORMANCE ANALYSIS

The frame length and frame rate are set to 25 ms and 10 ms (frame shift between two consecutive frames), respectively, to obtain acoustic features (LPC/MFCC/PLP) from an input speech. These features comprised of 39 dimensional features vector.

The next steps are training the LSTM model using the recorded training set of data for Bangla speech. After the training process the models will be ready for testing or recognizing the taste speech samples. designing an accurate continuous speech recognizer, sentence correct rate (SCR) for data set are evaluated. Three popular feature extraction methods, namely, LPC, MFCC and

PLP has been tried using the following combination of training and testing Bangla speech samples for each case.

Set A- Train 10000 male female, Test 1000 female
Set B- Train 5000 female, Test 1000 female
Set C- Train 5000 male, Test 1000 female
Set D- Train 10000 male female, Test 1000 male
Set E- Train 5000 female, Test 1000 male
Set F- Train 5000 male, Test 1000 male

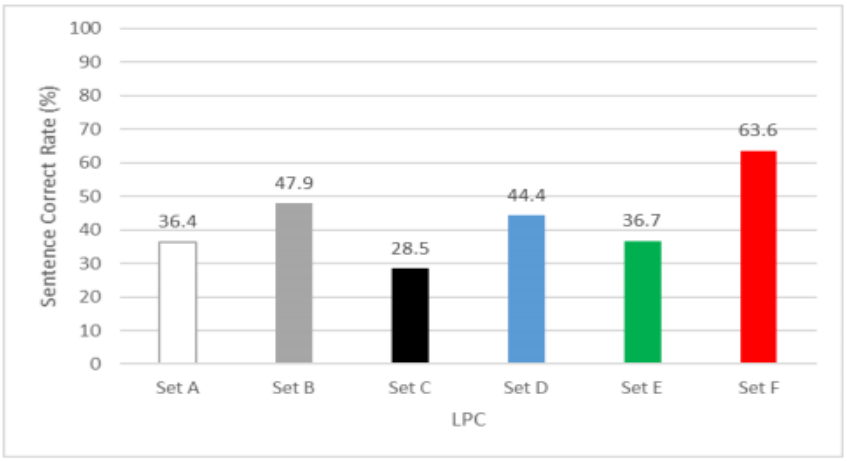

Figure 5. Sentence correct rate using the LPC as feature extraction method.

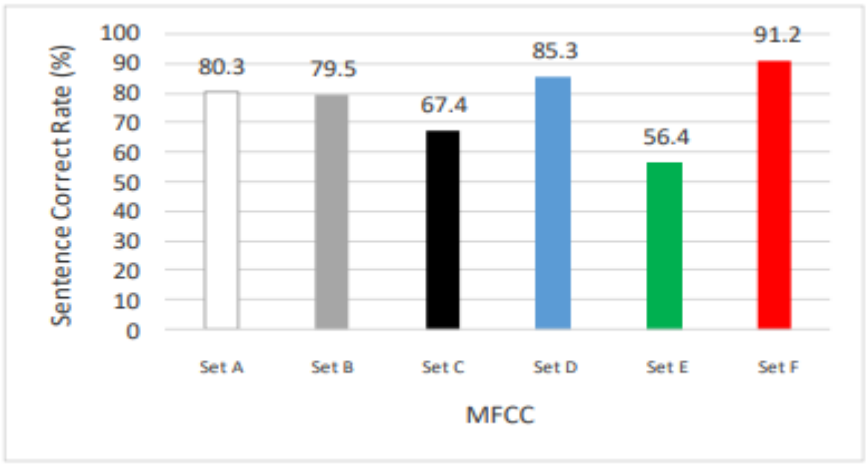

Figure 6. Sentence correct rate using the MFCC as feature extraction method.

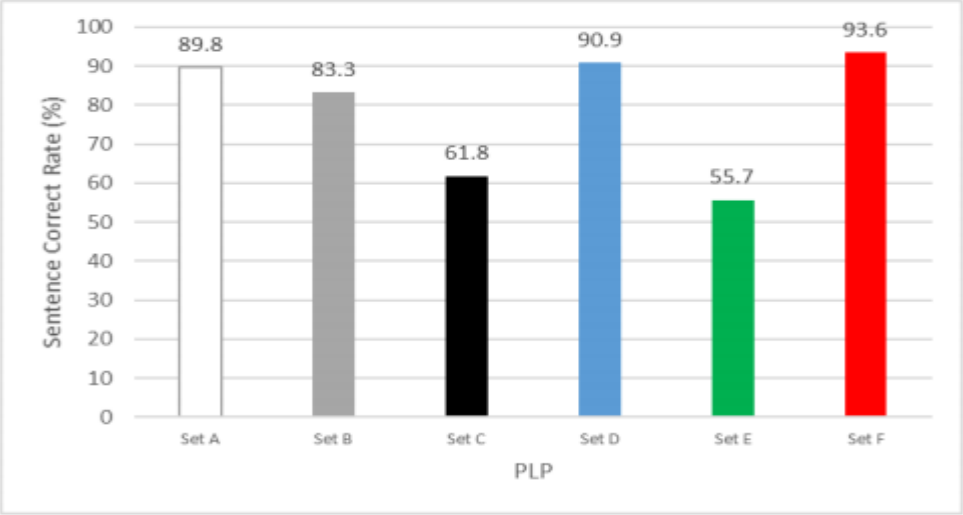

Figure 7. Sentence correct rate using the PLP as feature extraction method.

Here sentence correct rate (SCR) is the percentage sentence-level accuracy based on a certain number of tests performed.

$$SCR = \frac{H}{N} \times 100\% \ldots (5)$$

Where H=Number of sentence correctly detected and N=Total number of tests.

6. CONCLUSION

From Figure 5, Figure 6 and Figure 7 it is clear that the overall performance of PLP is better than that of MFCC. Here LPC seems to be the worst performer. However, it is clear that both PLP and MFCC and even LPC has a bias towards male speakers. But if we set aside the cross gender effect of PLP, it shows an overall balanced execution. To minimize the gender effect, the detector should be trained with a large set of both male and female data.

REFERENCES

[1] Kishore, S., Black, A., Kumar, R., and Sangal, R. "Experiments with Unit Selection Speech Databases for Indian Languages," National Seminar on Language Technology Tools: Implementation of Telugu, Hyderabad, India, October 2003.

[2] Karim R., Rahman M. S., and Iqbal M. Z., "Recognition of Spoken Letters in Bangla," 5th International Conference on Computer and Information Technology (ICCIT02), Dhaka, Bangladesh, 2002.

[3] Hassan M. R., Nath B., and Bhuiyan M. A.. "Bengali Phoneme Recognition: A New Approach," 6th International Conference on Computer and Information Technology (ICCIT03), Dhaka, Bangladesh, 2003.

[4] Rahman K. J., Hossain M. A., Das D., Islam T., and Ali M. G., "Continuous Bangla speech recognition system," 6th International Conference on Computer and Information Technology (ICCIT03), Dhaka, Bangladesh, 2003.

[5] Hossain S. A., Rahman M. L., Ahmed F., and Dewan M., "Bangla speech synthesis, analysis, and recognition: an overview," National Conference on Computer Processing of Bangla (NCCPB04), Dhaka, 2004.

[6] Hasnat M. A., Mowla J., and Khan M., "Isolated and Continuous Bangla Speech Recognition: Implementation Performance and Application Perspective," International Symposium on Natural Language Processing (SNLP), Hanoi, Vietnam, December 2007.
[7] Kotwal M. R A., Bonik M., Eity Q. N., Huda M. N., Muhammad G., Alotaibi Y. A., "Bangla Phoneme Recognition for ASR Using Multilayer Neural Network," International Conference on Computer and Information Technology (ICCIT10), Dhaka, Bangladesh, December, 2010.
[8] Kotwal M. R. A., Hassan F., Daud S. I., Alam M. S., Ahmed F., Huda M. N., "Gender effects suppression in Bangla ASR by Designing Multiple HMM-based Classifiers," CICN 2011, Gwalior, India, October 2011.
[9] Muhammad G., Alotaibi Y. A., Huda M. N., "Automatic Speech Recognition for Bangla Digits," International Conference on Computer and Information Technology (ICCIT09), Dhaka, Bangladesh, December 2009.
[10] Alam, F., Habib, S. M., Sultana, D. A., Khan M., (2010). "Development of Annotated Bangla Speech Corpora." BRAC University Institutional Repository. Retrieved June 28, 2012, from http://dspace.bracu.ac.bd/handle//10361//633.
[11] Wikipedia (2020). Audacity (audio editor). Retrieved from http://en.wikipedia.org/wiki/Audacity
[12] Wikipedia (2020). WAV. Retrieved from http://en.wikipedia.org/wiki/WAV
[13] Bengali alphabet Omniglot Retrieved from http://www.omniglot.com/writing/bengali.htm
[14] Hassan F., Kotwal M. R A., Alam M. S., and Huda M. N., "Gender Independent Bangla Automatic Speech Recognition," IEEE/IAPR Internationaol Conference on Informatics, Electronics and Vision (ICIEV) 2012, May 2012, Dhaka, Bangladesh .
[15] Atal, B. S., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." The Journal of The Acoustical Society of America 47 (1970) 65.
[16] Davis, S. B and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, no. 4, pp. 357 – 366, August 1980.
[17] Sepp Hochreiter and Jurgen Schmidhuber, "Long short-term emory", Neural Computation, vol.9,no.8,pp.1735 780,Nov.1997
[18] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. "State-of-the art speech recognition with sequence-to-sequence models". In Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference, pages 4774–4778. IEEE, 2018.
[19] Hasim Sak, Andrew Senior, and Francoise Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition", ArXiv e-prints, Feb.2014.
[20] Zhehuai Chen, Yimeng Zhuang, Yanmin Qian, Kai Yu, Zhehuai Chen, Yimeng Zhuang, Yanmin Qian, Kai Yu, Kai Yu, Yimeng Zhuang, et al. "Phone synchronous speech recognition with CTC lattices" IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 25(1): 90–101, 2017

AUTHOR

Nahyan Al Mahmud was born in Dhaka, Bangladesh, in 1987. He graduated from Electrical and Electronic Engineering department of Ahsanullah University of Science and Technology (AUST), Dhaka in 2008. Mr. Mahmud has completed the MSc program (EEE) from Bangladesh University of Engineering & Technology (BUET), Dhaka. Currently he is working as an Assistant Professor of EEE Department in AUST. His research interests include system and signal processing, analysis and design.