International Journal of Applied Information Systems (IJAIS) ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12 - No. 24, October 2019 - www.ijais.org

Neural Network on the Performance of Bangla Automatic Speech Recognition

Qamrun Nahar Eity
Dept. of CSE
Ahsanullah University
0f Science & Technology
Dhaka, Bangladesh

Md. Khairul Hasan
Dept. of CSE
Ahsanullah University
0f Science & Technology
Dhaka, Bangladesh

G. M. Monjur Morshed Mrida
Dept. of CSE
United International University
Dhaka, Bangladesh

Mohammad Nurul Huda, PhD
Dept. of CSE
United International University
Dhaka, Bangladesh

ABSTRACT

In this paper, the performance of different Bangla (widely used as Bengali) Automatic Speech Recognition (ASR) systems based on local features (LFs) to observe the effects of multilayer neural network (MLN) on it, is evaluated. These ASR systems use 3000 sentences uttered by 30 speakers from a wide area of Bangladesh, where Bangla is used as a native language. In the experiments, at first LFs are extracted from the input speech and these LFs are inputed into a multilayer neural network (MLN) for obtaining phoneme probabilities for all the Bengali phonemes considered in this study. Then, these phoneme probabilities are modified by taking logarithm or normal values, and either of these values are inputted to the hidden Markov model (HMM) based classifier to obtain word corrrect rate (WCR), word accuracy(WA) and sentence correct rate (SCR). From this study, it is observed that the ASR method which incorporates an MLN in its arechitecture improves the word recognition accuracy with fewer components in HMMs.

Keywords

local features; multi layer neural network; boost up; logarithm; normalization; hidden Markov model; automatic speech recognition.

1. INTRODUCTION

More than 220 million people speak in Bangla, which is one of the largely spoken languages in the world, as their native language, but the automatic speech recognition (ASR) researches using this language are few though it is ranked seventh based on the number of speakers [1].Because of the lackings of the speech corpora it is very difficult to research using the Bangla language,but Bangla speech corpus to build a Bangla text to speech system [2] are developed for Indian Bangla Language (West Bengal and Kolkata as its capital) by eliminating this problem at a limited extent. Though most of the natives of Bangla (more than two thirds) reside in Bangladesh, where it is the official language, but it has no speech corpus for its own accent. Although the written characters of standard Bangla in both the countries are same, there are some sound that are produced variably in different pronunciations of standard Bangla,in addition to the myriad of phonological variations in non-standard dialects [3].Therefore,a large scale speech corpus for Bangla language using Bangladeshi accent is badly needed for research purpose.

Some Bangla ASR systems were constructed in [4]-[11],but they concentrate on a simple recognition task on a very small database, or simply on the frequency distributions of different vowels and consonants. Here, an ASR system for Bangla in a large scale is built. For this purpose,a medium size Bangla speech corpus comprises native speakers covering almost all the major cities of Bangladesh is developed .Then local features (LFs) extracted from the input speech are inserted into multilayer neural network (MLN), and finally extracted features, phoneme probabilities, with some modifications (logarithm and normalization) are inserted into the hidden Markov model (HMM) based classifier for obtaining the word recognition performance. For evaluating Bangla word correct rate (WCR), word accuracy (WA) and sentence correct rate (SCR), there is designed two experiments (a) LF75+HMM and (b)LF75+MLN+HMM [With Boasting].

The paper is organized as follows. Section 2 briefly describes approximate Bangla phonemes with its corresponding phonetic symbols; Section 3 explains about Bangla speech corpus; Section 4 provides a brief description about the extraction of Local Feature, while Section 5 describes the different ASR systems. Section 6 explains the speech corpus and experimental setup. Section 7 explicates the experimental results and discussion, and finally, Section 8 draws some conclusions and remarks on the future works.

2. BANGLA PHONEMES WITH PHONETIC SYMBOLS

Phonetic inventory of Bangla consists of 8 short vowels, excluding long vowels, and 29 consonants. Native Bangla words do not allow initial consonant clusters: the maximum syllable structure is CVC (i.e. one vowel flanked by a consonant on each side) [12]. Sanskrit words borrowed into Bangla possess a wide range of clusters, expanding the maximum syllable structure to CCCVC. English or other foreign borrowings add even more cluster types into the Bangla inventory.

3. BANGLA SPEECH CORPUS

At present, a real problem to do experiment on Bangla phoneme ASR is the lack of proper Bangla speech corpus. In fact, such a corpus is not available or at least not referenced in any of the existing literature. Therefore,a medium size Bangla speech corpus, which is described below is developed.

Hundred sentences from the Bengali newspaper Prothom Alo [13] are uttered by 30 male speakers of different regions of Bangladesh. These sentences (30x100) are used for corpus. All of the speakers are Bangladeshi nationals and native speakers of Bangla. The age of the speakers ranges from 20 to 40 years.Speakers have chosen from a wide area of Bangladesh: Dhaka (central region), Comilla Noakhali (East region), Rajshahi (West region), Dinajpur Rangpur (North-West region), Khulna (South-West region), Mymensingh and Sylhet (North-East region). Though all of them speak in standard Bangla, they are not free from their regional accent.

Recording was done in a quiet room located at United International University (UIU), Dhaka, Bangladesh. A desktop was used to record the voices using a head mounted close-talking microphone. Recording of the voice was done in a place, where ceiling fan and air conditioner were switched on and some low level street or corridor noise could be heard.

Jet Audio 7.1.1.3101 software was used to record the voices. The speech was sampled at 16 kHz and quantized to 16 bit stereo coding without any compression and no filter is used on the recorded voice.

4. LOCAL FEATURE EXTRACTION

At an acoustic feature extraction stage in [14], firstly, input speech is converted into local features (LFs) that represent a variation in spectrum along time and frequency axes. Two LFs are first extracted by applying three point linear regressions (LRs) along the time t and frequency f axes on a time spectrum pattern respectively. Fig. 1 exhibits an example of LFs for an input utterance. After compressing these two LFs with 24 dimensions into LFs with 12 dimensions using discrete cosine transform (DCT), a 25-dimensional ($12\Delta t, 12\Delta f and \Delta P$, where P stands for log power of raw speech signal) feature vector named LF is extracted (see fig. 2).

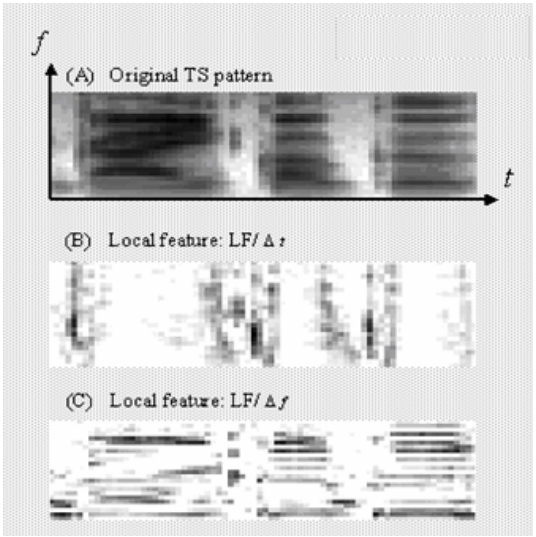

Fig. 1. Examples of LFs

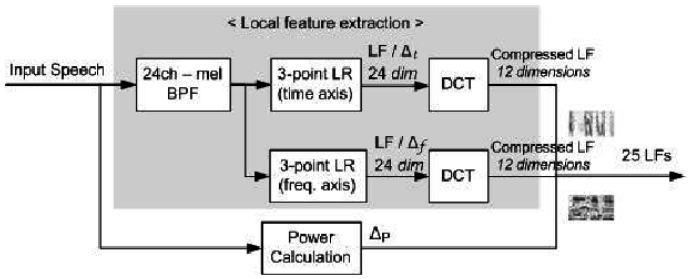

Fig. 2. LFs Extraction Procedure

5. ASR SYSTEMS

5.1 ASR Without MLN

LFs of preceding (t-3)-th, succeeding (t+3)-th frames with the current t-th frame that are then inputted directly to HMM based classifier [fig. 3].

5.2 ASR With MLN

In this method shown in fig. 4, LFs are inputted to the MLN with four layers, including three hidden layers. The MLN has 53 output units (all phonemes including sp and sil) of phoneme probabilities for the current frame t. The three hidden layers consist of 400, 200 and 100 units, respectively. The MLN is trained by using the standard back-propagation algorithm for 200 epochs. After extracting phoneme probabilities, logarithm (10-base) and normalization is applied on these phoneme probabilities (53) [fig. 4]. For normalization the following formula is used.

$$Y_i = \frac{X_i}{\sqrt{\sum_{j=1}^{j=53}(X_j)}} \quad (1)$$

The MLN is boosted up by adding the input data of non-matched output with the original trained data after 200 epochs. Then train the MLN several times with the same procedure up to 300 epochs.

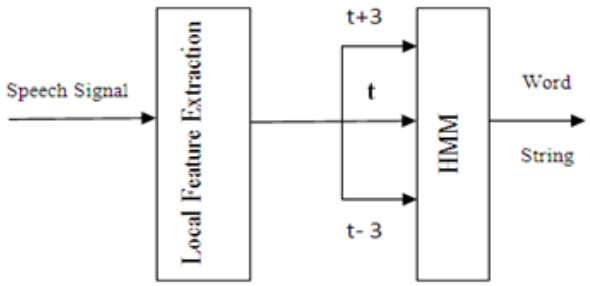

Fig. 3. ASR without MLN

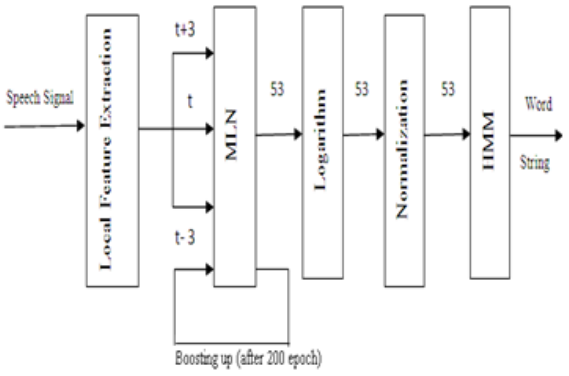

Fig. 4. ASR with MLN

Table 1. Sentence and Word accuracy for data sets 1,2,3 and 4 for Mixture-1

Mixture-1			
Data Base		LF75	LF75+MNN+HMM(With Boasting)
Set:1	SENT	90	95.47
	WORD	90.88	95.77
Set:2	SENT	86.8	94.93
	WORD	88.24	95.19
Set:3	SENT	89.07	94.53
	WORD	90.15	94.55
Set:4	SENT	91.47	96.27
	WORD	92.63	96.58

6. EXPERIMENTS

6.1 Speech Database

Here, speech corpus consists of 3000 sentences is used for training and testing. These data sets are divided into 4 segments using the cross validation strategy like below.

Set 1: Train: 1 to 2250; Test: 2251 to 3000
Set 2: Train: 751 to 3000; Test: 1 to 750
Set 3: Train: 1 to 750 1501 to 3000; Test: 751 to 1500
Set 4: Train: 1 to 1500 2251 to 3000; Test: 1501 to 2250

6.2 Experimental Setup

For all the methods, mixture components for HMM are set to 1, 2, 4 and 8. This is common for all data sets.

Sigmoid function is used as non-linear function for MLN are 90.88% and 95.77%. Table 2, 3, 4 show the same but for mixture 2, 4 and 8 respectively. Table 5 shows the average of mixture 1, 2, 4 and 8 for sentence and word accuracy of data sets 1, 2, 3 and 4. Table 6 shows average of sentence and word accuracy on average of mixture 1, 2, 4 and 8 for sentence and word accuracy of data sets 1, 2, 3 and 4, where we found, the sentence accuracy for LF75+HMM and LF75+MLN+HMM [With Boasting] are 87.41% and 95.51% and the word accuracy are 88.61% and 95.77%. Here, LF75 method gives poor performance and with MLN the performance increases.

Figures 5-10 show the chart representation of tables 1-6 respectively.

7. EXPERIMENTAL RESULTS AND DISCUSSION

Table 1 shows the sentence and word accuracies for all data sets using the mixture components one, two, three and four for the investigated methods, where, in data set 1 the sentence accuracy for LF75+HMM and LF75+MLN+HMM [With Boasting] are 90%,and 95.47% and the word accuracy are 90.88% and 95.77%. Table 2, 3, 4 show the same but for mixture 2, 4 and 8 respectively. Table 5 shows the average of mixture 1, 2, 4 and 8 for sentence and word accuracy of data sets 1, 2, 3 and 4. Table 6 shows average of sentence and word accuracy on average of mixture 1, 2, 4 and 8 for sentence and word accuracy of data sets 1, 2, 3 and 4, where we found, the sentence accuracy for LF75+HMM and LF75+MLN+HMM [With Boasting] are 87.41% and 95.51% and the word accuracy are 88.61% and 95.77%. Here, observation done that LF75 method gives poor performance and with MLN the performance increases.

Figures 5-10 show the chart representation of tables 1-6 respectively.

Table 2. Sentence and Word accuracy for data sets 1,2,3 and 4 for Mixture-2

Mixture-2			
Data Base		LF75	LF75+MNN+HMM(With Boasting)
Set:1	SENT	90.67	95.07
	WORD	90.61	95.42
Set:2	SENT	87.2	95.2
	WORD	88.93	95.44
Set:3	SENT	88	95.73
	WORD	88.8	96.01
Set:4	SENT	91.2	96.67
	WORD	92.63	96.92

Table 3. Sentence and Word accuracy for data sets 1,2,3 and 4 for Mixture-4

Mixture-4			
Data Base		LF75	LF75+MNN+HMM(With Boasting)
Set:1	SENT	90.53	94.4
	WORD	91.54	94.76
Set:2	SENT	86	94.8
	WORD	87.76	94.92
Set:3	SENT	85.2	96
	WORD	86.33	96.25
Set:4	SENT	88.53	96.93
	WORD	90.04	97.23

Table 4. Sentence and Word accuracy for data sets 1,2,3 and 4 for Mixture-8

Mixture-8			
Data Base		LF75	LF75+MNN+HMM(With Boasting)
Set:1	SENT	84.53	95.6
	WORD	85.15	95.98
Set:2	SENT	84.27	95.2
	WORD	86	95.37
Set:3	SENT	81.47	95.07
	WORD	82.83	95.32
Set:4	SENT	83.6	96.27
	WORD	85.17	96.54

Table 5. Average of mixture 1, 2, 4 and 8 for Sentence and Word accuracy of data sets 1, 2, 3 and 4

Average			
Data Base		LF75	LF75+MNN+HMM(With Boasting)
Set:1	SENT	88.9325	95.135
	WORD	89.545	95.4825
Set:2	SENT	86.0675	95.0325
	WORD	87.7325	95.23
Set:3	SENT	85.935	95.3325
	WORD	87.0275	95.5325
Set:4	SENT	88.7	96.535
	WORD	90.1175	96.8175

Table 6. Average of Sentence and Word accuracy on average of mixture 1, 2, 4 and 8 for Sentence and Word accuracy of data sets 1, 2, 3 and 4.

		LF75	LF75+MNN+HMM(With Boasting)
Set:1	SENT	87.40875	95.50875
	WORD	88.605625	95.765625

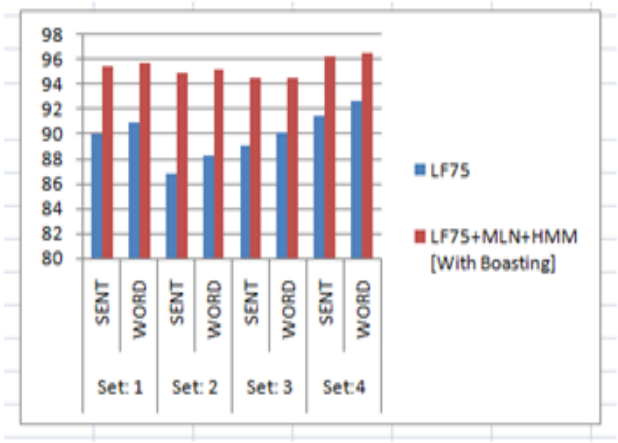

Fig. 5. Sentence and Word accuracy for data set 1, 2, 3 and 4 in Mixture-1.

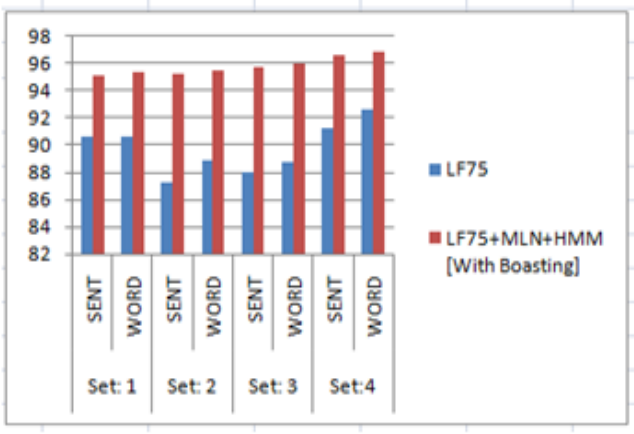

Fig. 6. Sentence and Word accuracy for data set 1, 2, 3 and 4 in Mixture-2.

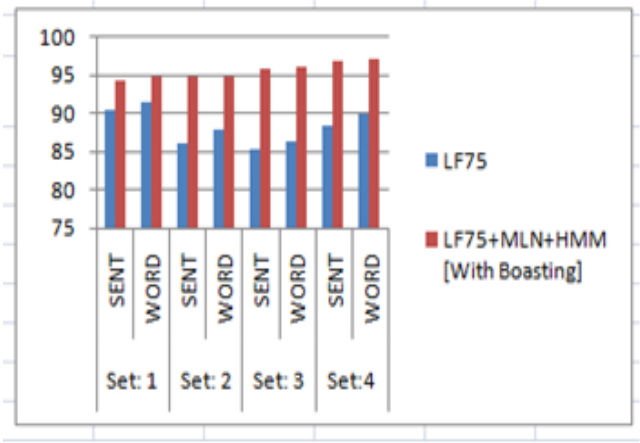

Fig. 7. Sentence and Word accuracy for data set 1, 2, 3 and 4 in Mixture-4.

8. CONCLUSION

This paper shows the effect of a neural network on the performance of an automatic speech recognition system, which incorporates local features as acoustic feature extraction procedure. The following conclusions are observed from this study.

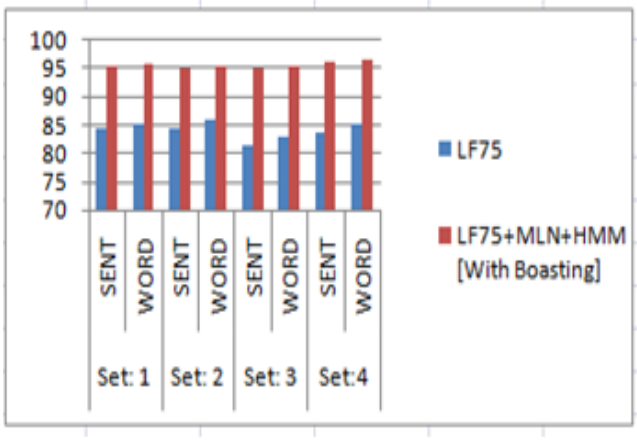

Fig. 8. Sentence and Word accuracy for data set 1, 2, 3 and 4 in Mixture-8.

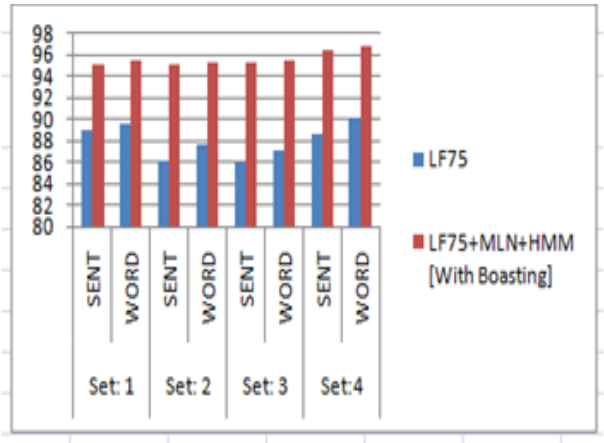

Fig. 9. Average of mixture 1, 2, 4 and 8 for Sentence and Word accuracy of data sets 1,2,3 and 4.

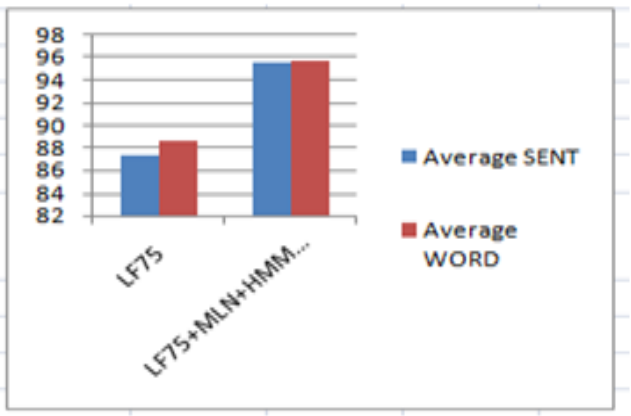

Fig. 10. Average Sentence and Word accuracy for method 1, 2, 3 and 4 on Average of mixture 1, 2, 4 and 8.

—The method incorporating MLN provides a higher performance with respect to the method that did not embed MLN.

—Fewer mixture components are required for LF-based methods.

The author would like to do further experiments for evaluating Bangla phoneme recognition incorporating a neural network in its architecture.

9. REFERENCES

[1] http://en.wikipedia.org/wiki/Listoflanguagesbytotalspeakers, Last accessed April 11, 2009.

[2] S. P. Kishore, A. W. Black, R. Kumar, and Rajeev Sangal, "Experiments with unit selection speech databases for Indian languages," Carnegie Mellon University.

[3] http://en.wikipedia.org/wiki/Bengaliphonology, Last accessed April 11, 2009.

[4] S. A. Hossain, M. L. Rahman, and F. Ahmed, Bangla vowel characterization based on analysis by synthesis, Proc. WASET, vol. 20, pp. 327-330, April 2007.

[5] M. A. Hasnat, J. Mowla, and Mumit Khan, " Isolated and Continuous Bangla Speech Recognition: Implementation Performance and application perspective, " in Proc. International Symposium on Natural Language Processing (SNLP), Hanoi, Vietnam, December 2007.

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

[7] A. K. M. M. Houque, "Bengali segmented speech recognition system," Undergraduate thesis, BRAC University, Bangladesh, May 2006.

[8] K. Roy, D. Das, and M. G. Ali, "Development of the speech recognition system using artificial neural network," in Proc. 5th International Conference on Computer and Information Technology (ICCIT02), Dhaka, Bangladesh, 2002

[9] M. R. Hassan, B. Nath, and M. A. Bhuiyan, "Bengali phoneme recognition: a new approach," in Proc. 6th International Conference on Computer and Information Technology (ICCIT03), Dhaka, Bangladesh, 2003.

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

[11] S. A. Hossain, M. L. Rahman, F. Ahmed, and M. Dewan, "Bangla speech synthesis, analysis, and recognition: an overview," in Proc. NCCPB, Dhaka, 2004.

[12] C. Masica, The Indo-Aryan Languages, Cambridge University Press, 1991.

[13] www.prothom-alo.com.

[14] M. R. A. Kotwal, F. Hassan, G. Md. M. Islam, M. Rakibuzzaman, M. M. Hasan, M. Banik, G. Muhammad and M. N. Huda, Bangla Phoneme Recognition for Different Acoustic Features, ICCAIE 2010, IEEE Explored, Kuala Lumpur, Malaysia, December, 2010.