DOI 10.1007/s10772-011-9095-3

MLN-based Bangla ASR using context sensitive triphone HMM

Foyzul Hassan · Mohammed Rokibul Alam Kotwal · Ghulam Muhammad · Mohammad Nurul Huda

Received: 2 February 2011 / Accepted: 13 June 2011 / Published online: 30 June 2011

Abstract Building a continuous speech recognizer for the Bangla (widely used as Bengali) language is a challenging task due to the unique inherent features of the language like long and short vowels and many instances of allophones. Stress and accent vary in spoken Bangla language from region to region. But in formal read Bangla speech, stress and accents are ignored. There are three approaches to continuous speech recognition (CSR) based on the sub-word unit viz. word, phoneme and syllable. Pronunciation of words and sentences are strictly governed by set of linguistic rules. Many attempts have been made to build continuous speech recognizers for Bangla for small and restricted tasks. However, medium and large vocabulary CSR for Bangla is relatively new and not explored. In this paper, the authors have attempted for building automatic speech recognition (ASR) method based on context sensitive triphone acoustic models. The method comprises three stages, where the first stage extracts phoneme probabilities from acoustic features using a multilayer neural network (MLN), the second stage designs triphone models to catch context of both sides and the final stage generates word strings based on triphone hidden Markov models (HMMs). The objective of this research is to build a medium vocabulary triphone based continuous speech recognizer for Bangla language. In this experimentation using Bangla speech corpus prepared by us, the recognizer provides higher word accuracy as well as word correct rate for trained and tested sentences with fewer mixture components in HMMs.

Keywords Continuous speech recognition · Triphone model · Bangla language · Multilayer neural network · Hidden Markov model

1 Introduction

Automatic speech recognition (ASR) deals with the decoding of an acoustic signal of a speech utterance into corresponding text transcription, such as words, phonemes or other language units. Even after years of extensive research and development, accuracy in ASR remains a challenge to researchers. There are number of well known factors which determine accuracy. The prominent factors are those that include variations in context, speakers and noise in the environment. Therefore, research in ASR has many open issues with respect to small or large vocabulary, isolated or continuous speech, speaker dependent or independent and environmental robustness.

ASR for western languages like English and Asian languages like Chinese are well matured. But similar research in Bangla (widely used as Bengali) languages is still in its infancy stage. Another major hurdle in ASR for the Bangla language is resource deficiency. Annotated speech corpora for training and testing the acoustic models are scarce. Recently there is a growing interest in ASR for Bangla language (Hossain et al. 2004, 2007; Hasnat et al. 2007; Karim et al. 2002; Houque 2006; Roy et al. 2002; Hassan

F. Hassan (✉) · M.R.A. Kotwal · M.N. Huda
Department of CSE, United International University, Dhaka, Bangladesh
e-mail: foyzul.hassan@gmail.com

M.R.A. Kotwal
e-mail: rokib39@gmail.com

M.N. Huda
e-mail: mnh@cse.uiu.ac.bd

G. Muhammad
Department of CE, College of CIS, King Saud University, Riyadh, Kingdom of Saudi Arabia
e-mail: gmd_babu@yahoo.com

Table 1 Bangla phonetic scheme in IPA ((Muhammad et al. 2009); http://en.wikipedia.org/wiki/Bengali_phonology)

(a) Vowel

	Front	Central	Back
Close	i		u
	i		u
Close-mid	e		o
	e		o
Open-mid	æ		ɔ
	ê		ô
Open		a	
		a	

(b) Consonants

		Labial	Dental/ Alveolar	Retroflex	Lamino-Postalveolar	Velar	Glottal
Nasal		m	n			ŋ	
		m	n			ng	
Plosive	voiceless	p	t̪	t	tʃ	k	
		p	t	ţ	ch	k	
	aspirated	p^{h1}	t̪^{h}	t^{h}	tʃ^{h2}	k^{h}	
		ph	th	ţh	chh	kh	
	voiced	b	d̪	ɖ	dʒ	g	
		b	d	đ	j	g	
	murmured	b^{ɦ}	d̪^{ɦ}	ɖ^{ɦ}	dʒ^{ɦ3}	g^{ɦ}	
		bh	dh	đh	jh	gh	
Fricative		f^{1}	s^{2}, z^{3}		ʃ^{2}		h
		f	s		sh		h
Approximant			I				
			I				
Rhotic			r^{4}	r̲^{4}			
			r	r̲			

et al. 2003; Rahman et al. 2003). Some have attempted to speech recognition for isolated word recognition in Bangla using Artificial Neural Network (ANN) (Roy et al. 2002; Hassan et al. 2003). Continuous Bangla speech recognition system is developed in Rahman et al. (2003), while Ming et al. (1998) presents a brief overview of Bangla speech synthesis and recognition. However, most of these researches have some problems: (i) deals with small scale speech corpus, (ii) directly inserts Mel frequency cepstral coefficients (MFCCs) to the HMM-based classifier and (iii) constructs triphone models (Thangarajan et al. 2008; Matousek et al. 2005; Dupont et al. 2005; Jurafsky et al. 2001; Ming et al. 1998) using MFCC features and consequently, better recognition performance is not obtained.

In this study, we have proposed a medium speech corpus based ASR system by designing triphone models constructed from phoneme probabilities instead of MFCC features. The proposed method comprises three stages, where the first stage extracts phoneme probabilities from acoustic features, MFCCs using a multilayer neural network (MLN), the second stage designs triphone models to catch context of both sides and the last stage outputs word strings based on triphone models using hidden Markov model based classifier. The specialties of this paper are (i) utilization of medium speech corpus instead of small one and (ii) triphone model construction using phoneme probabilities extracted from MFCCs instead of using direct MFCCs.

This paper is organized into eight sections including the introduction. Section 2 briefly describes approximate Bangla phonemes with its corresponding phonetic symbols. Sections 3 and 4 explain about the preparation of Bangla speech corpus and the modeling of triphones, respectively. Section 5 describes the system configuration of the conventional and the proposed methods. Section 6 provides an overview of experimental setup. Section 7 shows the results of the experimentation and their analysis, and Sect. 8 gives the conclusion of the paper and remarks on future works.

2 Bangla phonemes schemeses

2.1 Bangla phonemes

The Phonetic inventory of Bangla consists of 14 vowels, including seven nasalized vowels, and 29 consonants (Muhammad et al. 2009; http://en.wikipedia.org/wiki/Bengali_phonology). An approximate phonetic scheme in IPA is given in Table 1. In Table 1(a), only the main 7 vowel sounds are shown, though there exists two more long counterpart of /i/ and /u/, denoted as /i:/ and /u:/, respectively. These two long vowels are seldom pronounced differently than their short counterparts in modern Bangla. There is controversy on the number of Bangla 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) (Masica 1991). Sanskrit words borrowed into Bangla possess a wide range of clusters, expanding the maximum syllable structure to CCCVC. English or other

Table 2 Some Bangla words with their orthographic transcriptions and IPA

Bangla Word	English Pronunciation	IPA	Our Symbol
আমরা	AAMRA	/a m r a/	/aa m r ax/
আচরণ	AACHORON	/a tʃ r n/	/aa ch ow r aa n/
আবেদন	ABEDON	/a b æ d̪ n/	/ax b ae d aa n/

foreign borrowings add even more cluster types into the Bangla inventory.

2.2 Bangla words

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

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, we develop a medium size Bangla speech corpus, which is described below.

Hundred sentences from the Bengali newspaper "Prothom Alo" (Daily Prothom Alo www.prothom-alo.com) are uttered by 30 male speakers of different regions of Bangladesh. These sentences (30 × 100) are used for training corpus (D1). On the other hand, different 100 sentences from the same newspaper uttered by 10 different male speakers (total 1000 sentences) are used as test corpus (D2). All of the speakers are Bangladeshi nationals and native speakers of Bangla. The age of the speakers ranges from 20 to 40 years. We have chosen the speakers 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. We record the voice in a place, where ceiling fan and air conditioner were switched on and some low level street or corridor noise could be heard.

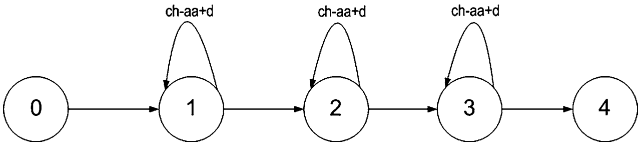


Fig. 1 Triphone model for the Bangla word চাঁদ

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. For this study, stereo coding has been used in order to reduce the redundancy in stereo coding signals. One can achieve significant signal gains in stereo coding which can be utilized to either boost the quality of the reconstructed signal or to lower the bit rate whiling keeping the signal quality constant with respect to the original coder.

4 Triphone design

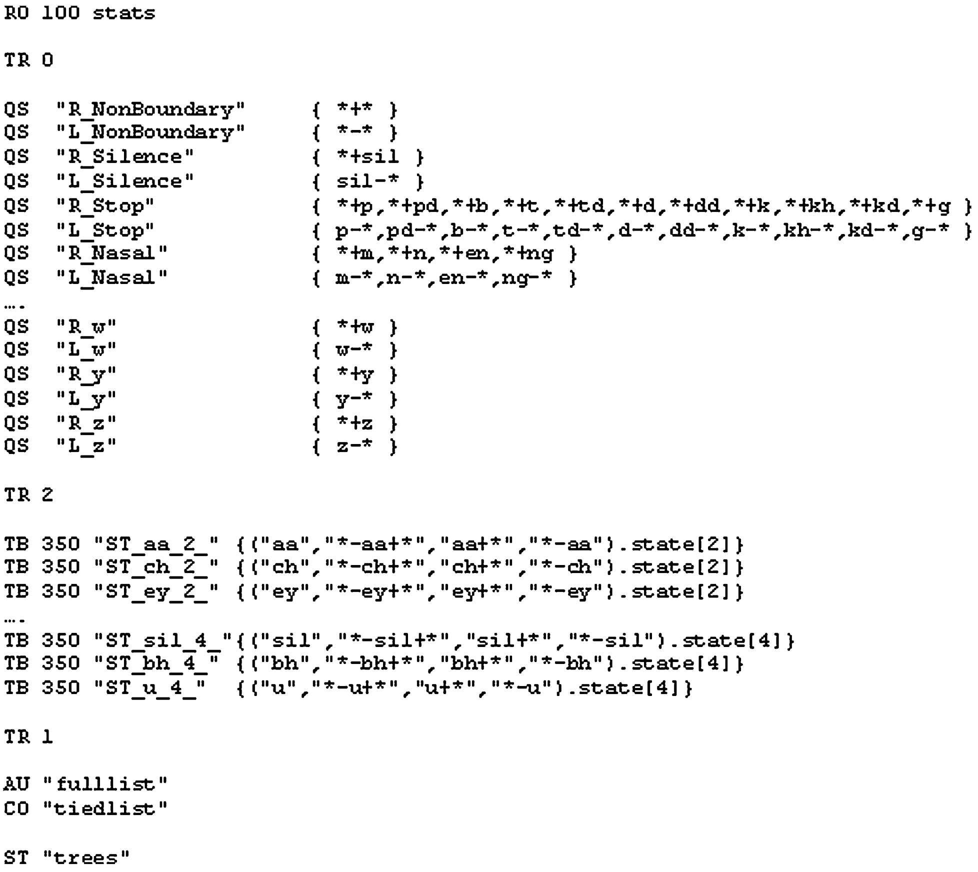
Figure 1 depicts a triphone model for the Bangla word চাঁদ (pronounced as 'chaad' and English meaning is 'Moon'). Here, five states three loops left-to-right HMM model is used.

Figure 2 shows the partial part of tree.hed, which contains the instructions regarding which contexts to examine for possible clustering, can be rather long and complex. It is noted that this script is only capable of creating the TB commands (decision tree clustering of states). The questions (QS) still need defining by the user. The entire script appropriate for clustering Bangla phone models is too long to show here in the text, however, its main components are given by the following fragments. Detail construction of these questions and their formats are given in HTK book http://www.cc.ntut.edu.tw/~enlab07/Documents/HTK/htkbook34.pdf.

5 System configuration

5.1 Conventional method

Traditional approach of ASR systems uses MFCC of 39 dimensions (12-MFCC, 12-ΔMFCC, 12-ΔΔMFCC, P, ΔP and ΔΔP, where P stands for raw log energy of the input speech signal) as feature vector to be fed into a HMM-based classifier and the system diagram is shown in Fig. 3. Parameters (mean and diagonal covariance of hidden Markov model of each phoneme) are estimated, from MFCC training data,



```
RO 100 stats

TR 0

QS  "R_NonBoundary"      { *+* }
QS  "L_NonBoundary"      { *-* }
QS  "R_Silence"          { *+sil }
QS  "L_Silence"          { sil-* }
QS  "R_Stop"             { *+p,*+pd,*+b,*+t,*+td,*+d,*+dd,*+k,*+kh,*+kd,*+g }
QS  "L_Stop"             { p-*,pd-*,b-*,t-*,td-*,d-*,dd-*,k-*,kh-*,kd-*,g-* }
QS  "R_Nasal"            { *+m,*+n,*+en,*+ng }
QS  "L_Nasal"            { m-*,n-*,en-*,ng-* }
….
QS  "R_w"                { *+w }
QS  "L_w"                { w-* }
QS  "R_y"                { *+y }
QS  "L_y"                { y-* }
QS  "R_z"                { *+z }
QS  "L_z"                { z-* }

TR 2

TB 350 "ST_aa_2_" {("aa","*-aa+*","aa+*","*-aa").state[2]}
TB 350 "ST_ch_2_" {("ch","*-ch+*","ch+*","*-ch").state[2]}
TB 350 "ST_ey_2_" {("ey","*-ey+*","ey+*","*-ey").state[2]}
….
TB 350 "ST_sil_4_"{("sil","*-sil+*","sil+*","*-sil").state[4]}
TB 350 "ST_bh_4_" {("bh","*-bh+*","bh+*","*-bh").state[4]}
TB 350 "ST_u_4_"  {("u","*-u+*","u+*","*-u").state[4]}

TR 1

AU "fulllist"
CO "tiedlist"

ST "trees"
```

Fig. 2 Fragments of tree.hed for Bangla phonemes

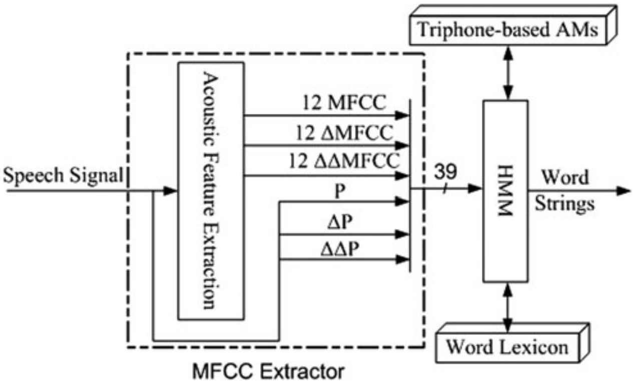


Fig. 3 Conventional approach of word recognizer

using Baum-Welch algorithm. For different mixture components, training data are clustered using the K-mean algorithm. Triphone models are configured using training data instead of monophone. During recognition phase, a most likely word for an input utterance is obtained using the Viterbi algorithm.

5.2 LF based method

At an acoustic feature extraction stage, 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. Figure 4 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 ΔP, where P stands for log power of raw speech signal) feature vector named LF is extracted (see Fig. 5). Recognition results using triphone model are found after inserting these 25-dimensional data vectors into HMM-based classifier, which is similar to the conventional approach.

5.3 Proposed method

Figure 6 shows the phoneme recognition method using MLN. At the acoustic feature extraction stage, input speech is converted into MFCCs of 39 dimensions that are input to an MLN with four layers, including three hidden layers, after combining preceding $(t-3)$-th and succeeding $(t+3)$-th frames with the current t-th frame. 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.

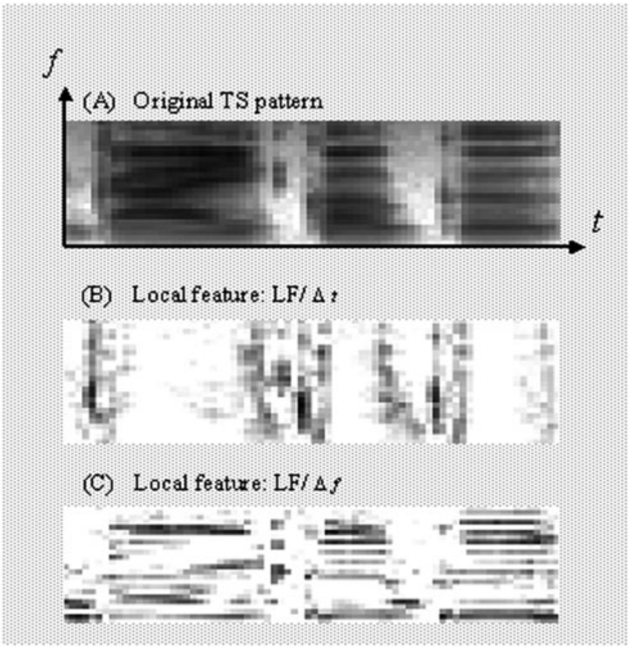


Fig. 4 Examples of LFs

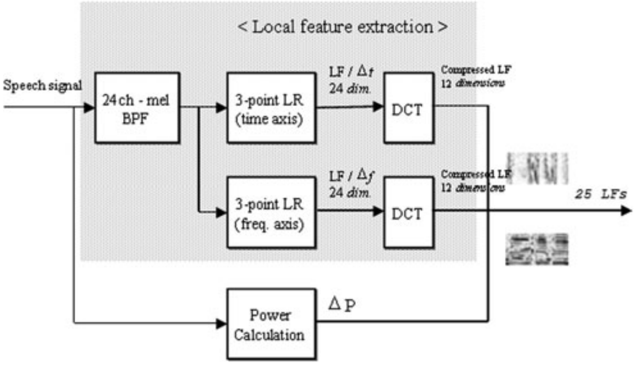


Fig. 5 LFs extraction procedure

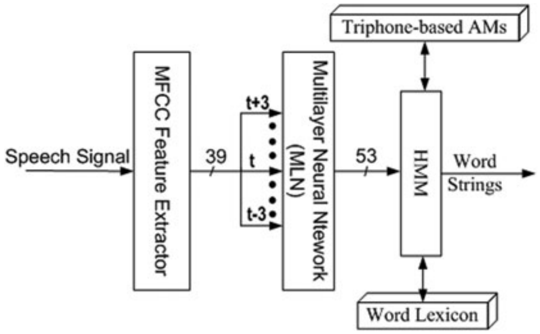


Fig. 6 Proposed word recognizer

These phoneme probabilities are inserted into the triphone based HMM to obtain more accurate word strings. The proposed method embeds short and long vowels, and some instances of allophones for designing triphone HMMs.

6 Experimental setup

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 (MFCCs) from an input speech. MFCC comprised of 39 dimensional features. LFs are a 25-dimensional vector consisting of 12 delta coefficients along time axis, 12 delta coefficients along frequency axis, and delta coefficient of log power of a raw speech signal (Nitta 1999).

For designing an accurate continuous word recognizer, word correct rate (WCR) and word accuracy (WA) for D2 data set are evaluated using an HMM-based classifier. The D1 data set is used to design Bangla triphones HMMs with five states, three loops, and left-to-right models. Input features for the classifier are 39 dimensional MFCC. In the HMMs, the output probabilities are represented in the form of Gaussian mixtures, and diagonal matrices are used. The mixture components are set to one, two, four and eight.

To obtain the WCR and WA we have designed the following experiments for D1 (close test) and D2 (open test) data sets:

(a) MFCC39+Triphone−HMM [Conventional].
(b) LF25+Triphone−HMM.
(c) MFCC39+MLN+Triphone−HMM [Proposed].

7 Experimental results and discussion

Figure 7 shows the comparison of WCR using D1 data set among the systems, MFCC39+Triphone−HMM, LF25+Triphone−HMM and MFCC39+MLN+Triphone−HMM. It is observed from the figure that the proposed system always provides higher WCR than the other method investigated at lower mixture components (One, Two and Four). For an example, at mixture component one, the proposed system exhibits 93.71% correct rate, while 87.13% and 90.08% WCRs are obtained by the methods, MFCC39+Triphone−HMM and LF25+Triphone−HMM, respectively. On the other hand, Fig. 8 gives corresponding WA for the methods investigated. It is also shown from this figure that similar types of results are obtained. These results exhibit the excellence of the proposed system over the other methods investigated. It is observed from the figure that the proposed method shows the highest word recognition performance over the methods, MFCC39+Triphone−HMM and LF25+Triphone−HMM at mixture component one. Figure 9 shows sentence correct rate (SCR) for the investigated methods using D1 data set. From the figure, it is shown that the proposed method always provides highest level correctness for the mixture components one, two and four. If we observe the Figs. 7, 8 and 9 for comparing the methods, (a) and (c), it provides strong evidence that the incorporation of

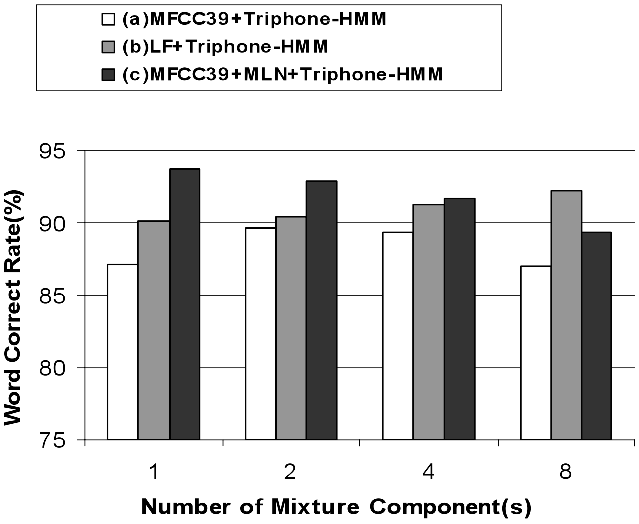


Fig. 7 Word correct rate for D1 data set

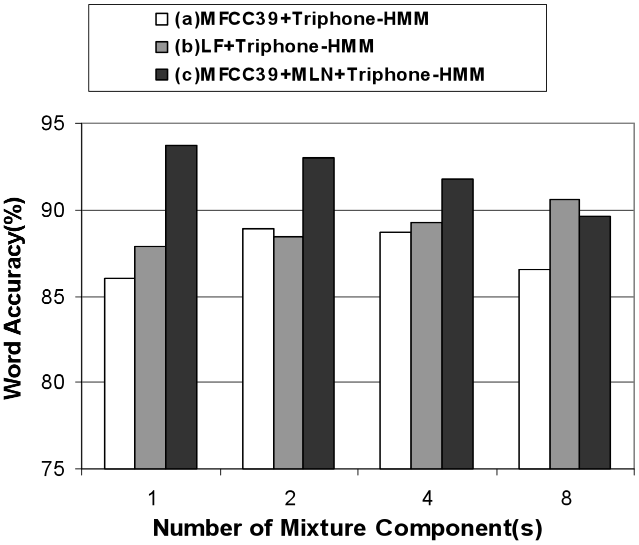


Fig. 8 Word accuracy for D1 data set

MLN with longer context input (Seven Frames: from "$t-3$" th frame to "$t+3$" th frame) has significant role for improving the performances. Moreover, the proposed method, (c) requires fewer mixture components to obtain better result than the other methods investigated.

Tables 3, 4 and 5 show the word recognition performance for the methods, MFCC39+Triphone−HMM, LF25+Triphone−HMM and MFCC39+MLN+Triphone−HMM, respectively using the D1 data set. It is observed from the tables that highest number of recognized words, H by the methods (c), (b) and (a) are 10824, 10658 and 10353 at mixture components one, eight and two respectively for the total number of training words, 11550. Again, the lowest number of total deletions for the methods (c), (b) and (a) are 217, 51 and 180 in mixture component one, four and one respectively, which indicates that the method (b) exhibits the lowest deletions among the all methods investigated. Moreover, the methods (c), (b) and (a) substitute 509, 839 and 989 words that represent lowest substitution at mixture component one, eight and four respectively. Finally, the least number of insertions for the corresponding methods, which measures the word accuracy, are found 13, 188 and 54 for the mixture component eight.

The sentence recognition performance for the methods (c), (b) and (a) are shown in Tables 6, 7 and 8 respectively using the D1 data set. The total numbers of correctly recognized sentences for the corresponding methods are 2813,

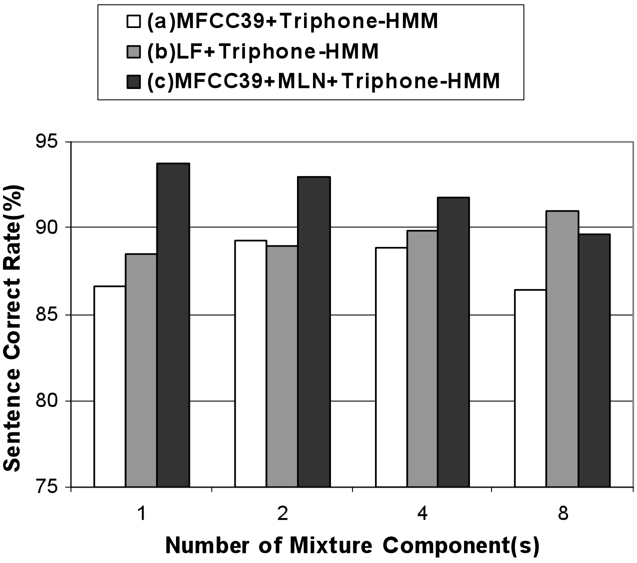


Fig. 9 Sentence correct rate for D1 data set

Table 3 Word recognition performance for MFCC39+TRIPHONE−HMM using D1 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	10063	10353	10320	10056
Deletion, D	180	183	241	391
Substitution, S	1307	1014	989	1103
Insertion, I	131	86	71	54
Total, *N*	11550	11550	11550	11550

Table 4 Word recognition performance for LF25+Triphone−HMM using D1 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	10404	10446	10537	10658
Deletion, D	60	68	51	53
Substitution, S	1086	1036	962	839
Insertion, I	246	233	225	188
Total, *N*	11550	11550	11550	11550

Table 5 Word recognition performance for MFCC39+MLN+Triphone−HMM using D1 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	10824	10726	10586	10327
Deletion, D	217	249	329	447
Substitution, S	509	575	635	776
Insertion, I	16	20	21	13
Total, *N*	11550	11550	11550	11550

Table 6 Sentence recognition performance for MFCC39+Triphone−HMM using D1 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	2598	2678	2667	2591
Substituted, S	402	322	333	409
Total, *N*	3000	3000	3000	3000

Table 7 Sentence recognition performance for LF25+Triphone−HMM using D1 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	2655	2669	2694	2729
Substituted, S	345	331	306	271
Total, *N*	3000	3000	3000	3000

Table 8 Sentence recognition performance for MFCC39+MLN+Triphone−HMM using D1 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	2813	2790	2753	2689
Substituted, S	187	210	247	311
Total, *N*	3000	3000	3000	3000

2729 and 2678 at mixture component one, eight and two out of 3000 training sentences, respectively. These recognized values indicate highest number of sentence correction rate among all the investigated mixture components. Our proposed method recognizes highest number of sentences because of having more context information over the other methods investigated.

WCRs using the D2 data set for the methods, MFCC39+Triphone−HMM, LF25+Triphone−HMM and MFCC39+MLN+Triphone−HMM systems are depicted in Fig. 10. From the figure, it is observed that the proposed system provides higher WCR than the other methods investigated at the lowest mixture components (One). The proposed system proclaims 91.91% correct rates at mixture component one, while 88.57% and 88.18% WCRs are obtained by the methods, MFCC39+Triphone−HMM and LF25+Triphone−HMM, accordingly. Beside the WCR, Fig. 11 illustrates WA for the corresponding methods investigated. These results exhibit that the proposed system provides better result than the other methods investigated. The proposed method produces highest recognition performance over the methods, MFCC39+Triphone−HMM and LF25+Triphone−HMM at mixture component one. Sentence correct rates (SCRs) for the investigated methods using the D2 data set are shown in Fig. 12. From the figure, it is clearly visible that the proposed method provides highest level correctness at mixture component one. Longer context window for the neural network input contain more information to resolve co-articulation effect and consequently, the

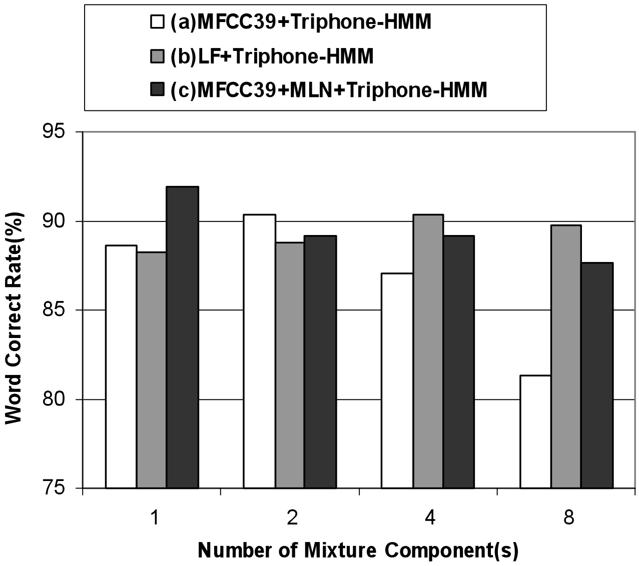


Fig. 10 Word correct rate for D2 data set

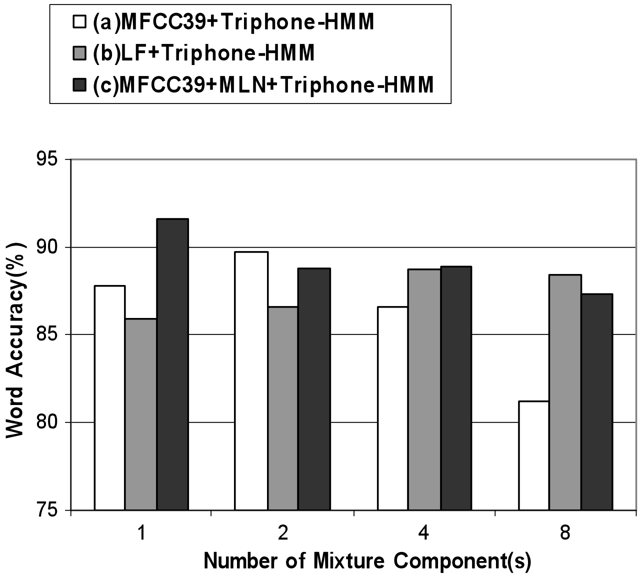


Fig. 11 Word accuracy for D2 data set

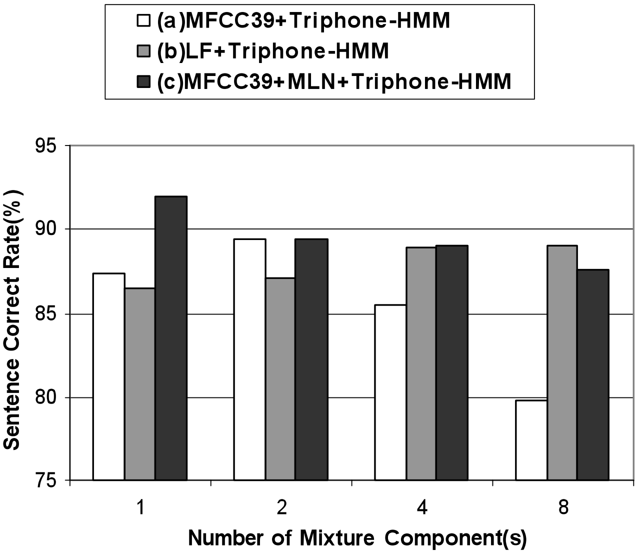


Fig. 12 Sentence correct rate for D2 data set

Table 9 Word recognition performance for MFCC39+ Triphone−HMM using D2 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	2914	2972	2865	2676
Deletion, D	88	74	104	181
Substitution, S	288	244	321	433
Insertion, I	24	21	17	5
Total, *N*	3290	3290	3290	3290

Table 10 Word recognition performance for MFCC39+ Triphone−HMM using D2 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	2901	2922	2973	2954
Deletion, D	43	38	33	45
Substitution, S	346	330	284	291
Insertion, I	74	73	56	47
Total, *N*	3290	3290	3290	3290

proposed system provides better result than the other methods investigated in fewer mixture component.

Word recognition performance for the methods, MFCC39 +Triphone−HMM, LF25+Triphone−HMM and the proposed, MFCC39+MLN+Triphone−HMM using the D2 data set are shown in Tables 9, 10 and 11. The proposed method outperformed the other methods for the correctly recognized word, H at mixture components one using D2 data set. The methods, (c), (b) and (a) provide the lowest number of deletions, 88, 33 and 74 at mixture components one, two and two, respectively. On the other hand, the lowest substitutions, 178, 284 and 244 at mixture components one, four and two are produced by the methods (c), (b) and (a), respectively. Finally, the methods, (c), (b) and (a) provide the lowest number of insertions, 9, 47 and 5 at mixture components one, eight and eight respectively.

Table 11 Word recognition performance for MFCC39+ MLN+Triphone−HMM using D2 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	3024	2935	2935	2882
Deletion, D	88	109	129	151
Substitution, S	178	246	235	257
Insertion, I	9	12	11	10
Total, *N*	3290	3290	3290	3290

Table 12 Sentence recognition performance for MFCC39+ Triphone−HMM using D2 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	874	894	855	798
Substituted, S	126	106	145	202
Total, *N*	1000	1000	1000	1000

Table 13 Sentence recognition performance for LF25+ Triphone−HMM using D2 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	865	871	889	890
Substituted, S	135	129	111	110
Total, *N*	1000	1000	1000	1000

Table 14 Sentence recognition performance for MFCC39+ MLN+Triphone−HMM using D2 data set

	Mix1	Mix2	Mix4	Mix8
Correctly Recognized, H	920	894	890	876
Substituted, S	80	106	110	124
Total, *N*	1000	1000	1000	1000

Tables 12, 13 and 14 depict sentence correct rates comparison among the methods, (a), (b) and (c) using the D2 data set. The highest number of correctly recognized sentences by the methods, (c), (b) and (a) are 920, 890 and 894 at mixture components one, eight and two out of 1000 test sentences, respectively. Our proposed method recognizes the highest number of sentences at lowest mixture component. The reason for these results is that the phoneme probabilities obtained by using the longer context window in MLN are used to design triphone HMM instead of acoustic features. This behavior is independent of any speech language, but related to neural network based model.

8 Conclusion

This paper has presented an MLN-based ASR system by designing context sensitive triphone HMMs. The following conclusions are drawn from this study:

(i) The proposed method shows highest level word correct rate, word accuracy and sentence correct rate at mixture component one.
(ii) Fewer mixture components in HMMs reduce required classification time in the proposed method.
(iii) The derived feature, phoneme probabilities obtained by using MLN in our proposed method over the acoustic feature, MFCC improves speech recognition performance significantly.

The author would like to do experiments using recurrent neural network (RNN) in near future. The authors also have intension to do experiments for resolving gender factor as a future research work using the proposed method.

References

Dupont, S., Ris, C., Couvreur, L., & Boite, J.-M. (2005). A study of implicit and explicit modeling of coarticulation and pronunciation variation. In *Proc. of InterSpeech'05*, Lisbon.

Hasnat, M. A., Mowla, J., & Khan, M. (2007). Isolated and continuous Bangla speech recognition: implementation performance and application perspective. In *Proc. international symposium on natural language processing (SNLP)*, Hanoi, Vietnam, December.

Hassan, M. R., Nath, B., & Bhuiyan, M. A. (2003). Bengali phoneme recognition: a new approach. In *Proc. 6th international conference on computer and information technology (ICCIT03)*, Dhaka, Bangladesh.

Hossain, S. A., Rahman, M. L., Ahmed, F., & Dewan, M. (2004). Bangla speech synthesis, analysis, and recognition: an overview. In *Proc. NCCPB*, Dhaka.

Hossain, S. A., Rahman, M. L., & Ahmed, F. (2007). Bangla vowel characterization based on analysis by synthesis. In *Proc. WASET* (Vol. 20, pp. 327–330).

Houque, A. K. M. M. (2006). *Bengali segmented speech recognition system*. Undergraduate thesis, BRAC University, Bangladesh, May 2006.

Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., & Sen, Z. (2001). What kind of pronunciation variation is hard for triphones to model. In *Proc. of ICASSP'01*, Salt Lake City, Utah.

Karim, R., Rahman, M. S., & Iqbal, M. Z. (2002). Recognition of spoken letters in Bangla. In *Proc. 5th international conference on computer and information technology (ICCIT02)*, Dhaka, Bangladesh.

Masica, C. (1991). *The Indo-Aryan languages*. Cambridge: Cambridge University Press.

Matoušek, J., Hanzlíček, Z., & Tihelka, D. (2005). Hybrid syllable/triphone speech synthesis. In *Proc. of InterSpeech'05*, Lisbon.

Ming, J. et al. (1998). Improved phone recognition using Bayesian triphone models. In *Proc. ICASSP'98*.

Muhammad, G., Alotaibi, Y. A., & Huda, M. N. (2009). Automatic speech recognition for Bangla digits. In *International conference on computer and information technology (ICCIT 2009)*, Dhaka, Bangladesh.

Nitta, T. (1999). Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA. In *Proc. ICASSP'99* (pp. 421–424).

Daily Prothom Alo. Online: www.prothom-alo.com.

Rahman, K. J., Hossain, M. A., Das, D., Islam, T., & Ali, M. G. (2003). Continuous Bangla speech recognition system. In *Proc. 6th international conference on computer and information technology (ICCIT03)*, Dhaka, Bangladesh.

Roy, K., Das, D., & Ali, M. G. (2002). Development of the speech recognition system using artificial neural network. In *Proc. 5th international conference on computer and information technology (ICCIT02)*, Dhaka, Bangladesh.

Thangarajan, R., Natarajan, A. M., & Selvam, M. (2008). Word and triphone based approaches in continuous speech recognition for Tamil language. In *WSEAS transactions on signal processing* (pp. 76–85).