Performance Evaluation of Bangla Word Recognition Using Different Acoustic Features

Nusrat Jahan Lisa[*1], Qamrun Nahar Eity[*2], Ghulam Muhammad[$]
Dr. Mohammad Nurul Huda[#1], Prof. Dr. Chowdhury Mofizur Rahman[#2]

[*] Department of Computer Science and Engineering Ahsanullah University of Science and Technology (AUST)
[$]Department of CE College of CIS, King Saud University Riyadh, Kingdom of Saudi Arabia
[#] Department of Computer Science and Engineering United International University

***Abstract*—** This paper describes a medium size Bangla speech corpus preparation and the comparison of the performances of different acoustic features for Bangla word recognition. A small number of speakers are use for most of the Bangla automatic speech recognition (ASR) system, but 40 speakers selected from a wide area of Bangladesh, where Bangla is used as a native language, are involved here. In the experiments, mel-frequency cepstral coefficients (MFCCs) and local features (LFs) are inputted the hidden Markov model (HMM) based classifiers for obtaining word recognition performance. From the experiments, it is shown that MFCC based method of 39 dimensions provides a higher word correct rate (WCR) than the other methods investigated. Moreover, a higher WCR is obtained by the MFCC39-based method with fewer mixture components in the HMM.

Keywords*—*mel-frequency cepstral coefficients, local features, hidden Markov model, automatic speech recognition, acoustic features

I. INTRODUCTION

Bangla (can also be termed as Bengali), which is largely spoken by the people all over the world, has been performed a very little research where many literatures in automatic speech recognition (ASR) systems are available for almost all the major spoken languages in the world. About 220 million or above people speak in Bangla as their native language. It is ranked seventh based on the number of speakers [1]. The lack of proper speech corpus is the major difficulty to research in Bangla ASR. Some efforts are made to develop Bangla speech corpus to build a Bangla text to speech system [2]. However, this effort is a part of developing speech databases for Indian Languages, where Bangla is one of the parts and it is spoken in the eastern area of India (West Bengal and Kolkata as its capital). But most of the natives of Bangla(more than two thirds) reside in Bangladesh, where it is the official language. 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, there is a need to do research on the main stream of Bangla, which is spoken in Bangladesh, ASR.

Manuscript received September 5, 2010
Manuscript revised September 20, 2010

Bangla ASR or Bangla speech processing research can be found in [4]-[11]. For example, Bangla vowel characterization is done in [4]; isolated and continuous Bangla speech recognition on a small dataset using hidden Markov models (HMMs) is described in [5]; recognition of Bangla phonemes by Artificial Neural Network (ANN) is reported in [8]-[9]. Continuous Bangla speech recognition system is developed in [10], while [11] presents a brief overview of Bangla speech synthesis and recognition. However, most of these works are mainly concentrated on simple recognition task on a very small database, or simply on the frequency distributions of different vowels and consonants.

We build an ASR system for Bangla word in a large scale for this study. We first develop a medium size (compared to the exiting size in Bangla ASR literature) Bangla speech corpus comprises of native speakers covering almost all the major cities of Bangladesh to achieve the goal. Then, melfrequency cepstral coefficients (MFCCs) and local features (LFs) are extracted from the input speech, then extracted features are inserted into MLN and finally the output of MLN are inserted into the hidden Markov model (HMM) based classifier for obtaining the word recognition performance. We have designed three experiments for evaluating Bangla word correct rate (WCR), (a) LF25+HMM, (b) MFCC38+HMM and (c) MFCC39+HMM.

The paper is arranged as follows. Section II briefly explains approximate Bangla phonemes with its corresponding phonetic symbols; Section III discusses about Bangla speech corpus; Section IV provides a brief description about MFCC-based and LF-based methods, while Section V describes experimental setup. Section VI explicates the experimental results and discussion, and finally, Section VII draws some conclusions and remarks on the future works.

II. PHONETIC SYMBOLS FOR BANGLA PHONEMES

Table I shows Bangla vowel phonemes with their corresponding International Phonetic Alphabet (IPA) and

my proposed symbols. Bangla phonetic inventory consists of 8 short vowels (A, Av, B, D, G, H, I, J), excluding long vowels (C, E) and 29 consonants. On the other hand, the consonants, which are used in Bangla language, are presented in Table III. Here, the Table exhibits the same items for consonants like as Table I. In the Table III, the pronunciation of /k/, /l/ and /m/ are same by considering the words ‡Kk (/keʃ/) , ‡gl (/meʃ/) and ‡Vm (/tʰeʃ/) respectively, which is shown in Fig. 1. Here the meaning of ‡Kk, ‡gl and ‡Vm are English language “hair”, “sheep” and “an insinuating remark” respectively. On the other hand, in the words Rvb (/dʒan/) and hvb (/dʒan/), there is no difference of pronunciation of /R/ and /h/ respectively that depicted in Fig. 2. Here the meaning of Rvb and hvb are English language “life” and “vehicle” respectively. Again, Fig. 3 shows that there is no difference of /Y/ and /b/ in the words cY (/pn/) and gb (/mn/) respectively. Here the meaning of cY and gb are English language “promise” and “mind” respectively. Moreover, phonemes /o/ and /p/ carry same pronunciation in the words Nvo (/gʰaʃ/) and Mvp (/gaʃ/) respectively, which is shown in the Fig. 4.

Initial consonant cluster is not allowed in the native Bangla: 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.

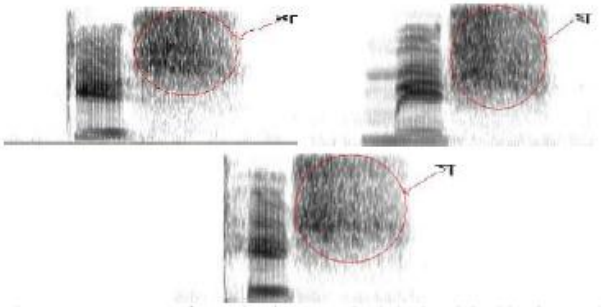

Fig. 1 Spectrogram of Bangla phonemes /শ/, /ম/ and /ত/ in the words কেশ (/keʃ/), মেষ (/meʃ/) and তেঁতা (/tʰeʃ/) respectively

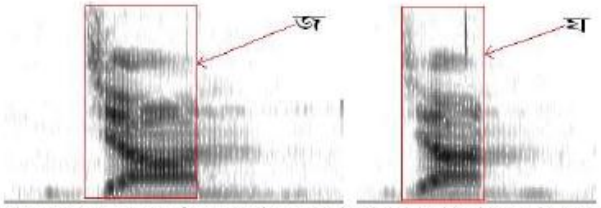

Fig. 2 Spectrogram of Bangla phonemes /জ/ and /য/ in the words জানা (/dʒan/) and যান (/dʒan/) respectively

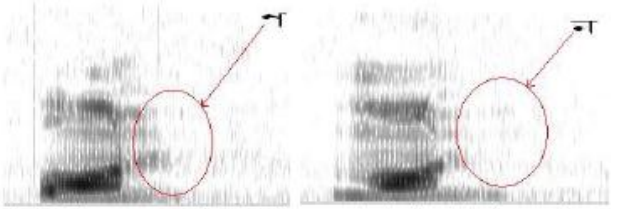

Fig. 3 Spectrogram of Bangla phonemes /প/ and /ন/ in the words পণ (/pn/) and মন (/mn/) respectively

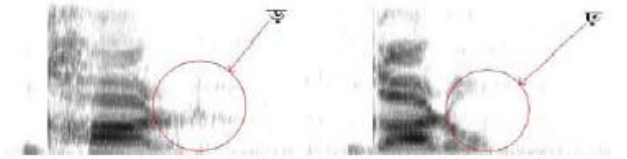

Fig. 4 Spectrogram of Bangla phonemes /ঙ/ and /ঢ/ in the words ঘাঙ (/gʰaʃ/) and গাঢ় (/gaʃ/) respectively

TABLE I
BANGLA VOWELS

Letter	IPA	Our Symbol
অ	/ɔ/ and /o/	a
আ	/a/	aa
ই	/i/	i
ঈ	/i/	i
উ	/u/	u
ঊ	/u/	u
এ	/e/ and /æ/	e
ঐ	/oj/	oi
ও	/o/	o
ঔ	/ow/	ou

TABLE III
BANGLA CONSONANTS

Letter	IPA	Our Symbol
ক	/k/	k
খ	/kʰ/	kh
গ	/g/	g
ঘ	/gʰ/	gh
ঙ	/ŋ/	ng
চ	/tʃ/	ch
ছ	/tʃʰ/	chh
জ	/dʒ/	j
ঝ	/dʒʰ/	jh
ট	/ʈ/	ta
ঠ	/ʈʰ/	tha
ড	/ɖ/	da
ঢ	/ɖʰ/	dha
ণ	/ɳ/	n
ত	/t̪/	t
থ	/t̪ʰ/	th
দ	/d̪/	d
ধ	/d̪ʰ/	dh
ন	/n/	n
প	/p/	p
ফ	/pʰ/	ph
ব	/b/	b
ভ	/bʰ/	bh
ম	/m/	m

Letter	IPA	Our Symbol
য	/dʒ/	j
র	/r/	r
ল	/l/	l
শ	/ʃ/ / /s/	s
ষ	/ʃ/	s
স	/ʃ/ / /s/	s
হ	/h/	h
ড়	/ɽ/	rh
ঢ়	/ɽ/	rh
য়	/e̯/ /-	y

III. BANGLA SPEECH CORPUS

Lack of proper Bangla speech corpus is the main problem to do experiment on Bangla word ASR. 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. From the Bengali newspaper "Prothom Alo" [13] hundred sentences are uttered by 30 speakers of different regions of Bangladesh. These sentences (30x100) are used for training corpus (D1). On the other hand, different 100 isolated words from the same newspaper uttered by 10 different female speakers (total 1000 isolated words) 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 Ahsanullah University of Science and Technology (AUST), 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.

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.

IV. SYSTEM CONFIGURATIONS

A. MFCC-based methods

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 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. 5. Parameters (mean and diagonal covariance of hidden Markov model of each phoneme) are estimated, from MFCC training data, using Baum-Welch algorithm. For different mixture components, training data are clustered using the K-mean algorithm. During recognition phase, a most likely word for an input utterance is obtained using the Forward algorithm. input utterance is obtained using the Forward algorithm. Another system based on MFCC of 38 dimensions (12-MFCC, 12-ΔMFCC, 12-ΔΔMFCC, P, ΔP and ΔΔP, where P stands for raw energy of the input speech signal) was designed.

B. LF-based method

At first input speech is converted into LFs at an acoustic feature extraction stage that represents a variation in spectrum along time and frequency axes [14]. Two LFs are first extracted by applying three-point linear regression (LR) along the time t and frequency f axes on a time spectrum pattern respectively. After compressing these two LFs with 24 dimensions into LFs with 12 dimensions using discrete cosine transform (DCT), a 25-dimensional (12 Δt, 12 Δf and ΔP, where P stands for log power of raw speech signal) feature vector named LF is extracted. Then, the extracted LFs are inserted into the HMM-based classifier for obtaining the output word. The procedure is shown in Fig. 6.

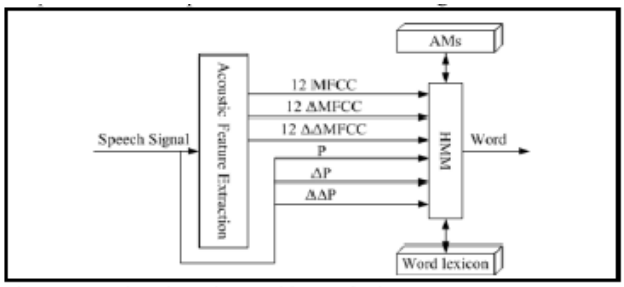

Fig. 5 MFCC-based word recognition methods

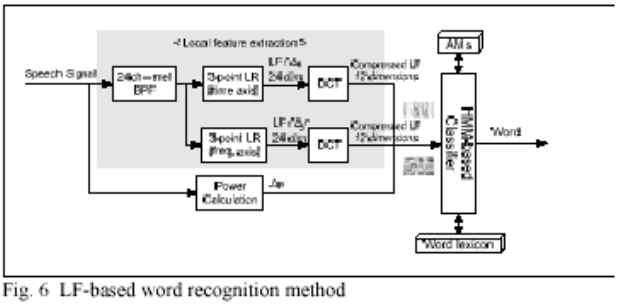

Fig. 6 LF-based word recognition method

V. EXPERIMENTAL SETUP

MFCC comprised of 38 and 39 dimensional. Acoustic feature vector 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 [14]. The frame length and frame rate are set to 25 ms and 10 ms, respectively, to obtain acoustic features (MFCCs or LFs) from an input speech.

For designing a word recognizer, WCR for D2 data set are evaluated using an HMM-based classifier. The D1 data set is used to design 39 Bangla monophone (8 vowels, 29 consonants, sp, sil) HMMs with five states, three loops, and left-to-right models. Input features for the classifier are 38 dimensional MFCC and 39 dimensional MFCC, and 25 dimensional LF for the MFCC-based and LF-based systems, respectively. 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 1, 2, 4, 8, 16 and 32. To obtain the WCR we have designed the following experiments

(a) LF25+HMM
(b) MFCC38 +HMM
(c)MFCC39 +HMM

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The comparison of WCR of test data set among LF25 +HMM, MFCC38+HMM and MFCC39+HMM systems is Shown in Table III. It is observed from the table that MFCC39-based system always provides higher WCR than the other method investigated. For an example, at mixture component 32, the MFCC39-based system exhibits 89.47% correct rate, while 78.91% and 85.55% WCRs are obtained by the methods LF25 +HMM and MFCC38 +HMM respectively.

TABLE IV
WORD CORRECT RATE FOR INVESTIGATED METHODS

Methods	Word Correct Rate (%)					
	Mix1	Mix2	Mix4	Mix8	Mix16	Mix32
LF25+MLN+HMM	50.50	62.38	69.90	75.74	76.04	78.91
MFCC38+MLN+HMM	49.80	64.75	72.45	78.39	83.85	85.55
MFCC39+MLN+HMM	55.25	70.20	79.70	85.56	86.14	89.47

Fig. 7 shows the comparison between MFCC38 and MFCC39. It is observed from the figure that, MFCC39 always exhibits lower word error rate over the method based on MFCC38. The reason for providing better result by MFCC39-based system is ΔΔP, where P stands for log power of raw speech signal along time axis.

Moreover, from the Table IV it is noted that MFCC39 requires fewer mixture components to obtain the approximately same numerical figure of correct rate provided by the method based on LF25 and MFCC38. For an example, Table V is given to indicate the computation time more specifically with the methods based on MFCC38 and MFCC39. We have measured the HMM time required by MFCC38 and MFCC39 by the formula mS^2T where m, S and T indicates number of mixture components, states and observation sequences respectively. For MFCC38, the required time is $32x5^2x200$ (=160K), while the corresponding time for the MFCC39 is $8x5^2x200$ (=40K) assuming number of observation sequence is 200 frames. Therefore, MFCC39 based method is faster than the method based on MFCC38.

TABLE V
Comparison of Time Complexity Between MFCC38 and MFCC39 Based Methods

	MFCC38	MFCC39
WCR=85.55	32 Mix	-
WCR=85.56	-	8 Mix
Required Multiplication	160K	40K

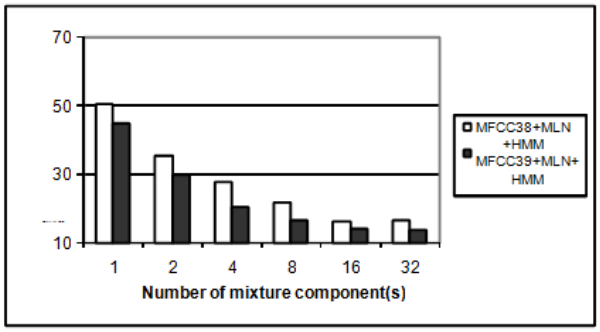

Fig. 7 Comparison between MFCC38 and MFCC39 based methods

VII. CONCLUSIONS

This paper compared performance of different acoustic features for Bangla word recognition and performed some experiments to obtain word recognition performance. A higher Bangla word correct rate for test data is also obtained by the MFCC39-based system. The author would like to do further experiments for obtaining Bangla word recognition performance after inserting all features into the neural network based systems.

REFERENCES

[1] http://en.wikipedia.org/wiki/List_of_languages_by_total_speakers, 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/Bengali_phonology, 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] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. ICASSP'99, pp.421-424, 1999.

Authors Profile:

Nusrat Jahan Lisa Received B.Sc. in Computer Science and Engineering degree from Ahsanullah University of Science and Technology (AUST) in 2007 and she is doing M.Sc. in Computer Science and Engineering in the United International University, Bangladesh . Currently She is the Lecturer of the Department of Computer Science and Engineering at the Ahsanullah University of Science and Technology (AUST), Dhaka, Bangladesh.

Qamrun Nahar Eity Received B.Sc. in Computer Science and Engineering degree from Ahsanullah University of Science and Technology (AUST) in 2008 and she is doing M.Sc. in Computer Science and Engineering in the United International University, Bangladesh. Currently she is the Lecturer of the Department of Computer Science and Engineering at the Ahsanullah University of Science and Technology (AUST), Dhaka, Bangladesh.

Ghulam Muhammad received Ph.D. Electronics and Information Engineering, Toyohashi University of Technology, Japan in March 2006. Currently he is the Assistant Professor, Department of Computer Engineering, College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia**.**

Dr. Mohammad Nurul Huda received Ph.D. (Electronics and Information Engineering, Toyohashi University of Technology,Japan) in 2008. Currently.he is the Associate Professor CSE, United International University, Dhaka, Bangladesh.

Prof. Dr. Chowdhury Mofizur Rahman received Ph.D. from Department of Computer Science, Tokyo Institute of Technology, Japan in 1996. Currently he is the Pro-Vice Chancellor, United International University, Dhaka, Bangladesh.