Proceedings of 2009 12th International Conference on Computer and Information Technology (ICCIT 2009)
21-23 December, 2009, Dhaka, Bangladesh

Automatic Speech Recognition for Bangla Digits

Ghulam Muhammad, Yousef A. Alotaibi, and Mohammad Nurul Huda†
College of Computer & Information Sciences, King Saud University, Riyadh, Saudi Arabia
†Computer Science and Engineering, United International University, Dhaka, Bangladesh
ghulam@ksu.edu.sa, yaalotaibi@ksu.edu.sa, mnh@cse.uiu.ac.bd

Abstract

In this paper, we introduce a system for Bangla digit automatic speech recognition (ASR). Though Bangla is one of the largely spoken languages in the world, only a few works on Bangla ASR can be found in the literature, especially on Bangladeshi accented Bangla. In this work, the corpus is collected from natives in Bangladesh. Mel-frequency cepstral coefficients (MFCCs) based features and hidden Markov model (HMM) based classifiers are used for recognition. Experimental results show comparatively high recognition performance (more than 95%) for first six digits (0 – 5) and low performance (less than 90%) for the next four digits (6 – 9). We notice two confused pairs of digits: one with '৬' (6) and '৯' (9), and the other with '৭' (7) and '৮' (8), in the experiments. We also find that different dialects in Bangladesh have a greater role on this confusion.

Keywords: Automatic speech recognition, Bangla digit, Bangla phoneme, hidden Markov model.

I. INTRODUCTION

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 works have been done in ASR for Bangla (can also be termed as Bengali), which is one of the largely spoken languages in the world. More than 215 million people speak in Bangla as their native language. It is ranked seventh based on the number of speakers [1]. A major difficulty to research in Bangla ASR is the lack of proper speech corpus. 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 is spoken in the eastern area of India (West Bengal). 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 sounds which are produced differently in different pronunciations of standard Bangla. Therefore, there is a need to do research on the main stream of Bangla, which is spoken in Bangladesh, ASR.

Some developments on Bangla speech processing or Bangla ASR can be found in [3]-[10]. For example, Bangla vowel characterization is done in [3]; isolated and continuous Bangla speech recognition on a small dataset using hidden Markov models (HMMs) is described in [4]; recognition of Bangla phonemes by Artificial Neural Network (ANN) is reported in [7]-[8]. Continuous Bangla speech recognition system is developed in [9], while [10] 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.

In this paper, we build an ASR system for Bangla digit. For this purpose, we first develop a medium size (compared to the exiting size in Bangla ASR literature) Bangla digit speech corpus comprises of native speakers covering almost all the major cities of Bangladesh. Then we build a Bangla digit ASR system using hidden Markov model toolkit (HTK) [11]. The results are investigated, and a group of confused Bangla digits in terms of ASR is analyzed. It is claimed that this is the first ever work done on Bangla digit ASR.

The paper is organized as follows. Section 2 briefly describes an approximate phonetic scheme for Bangla digits; Section 3 explains about Bangla digit speech corpus; Section 4 gives experimental setup, results and discussion on Bangla digit ASR. Finally, Section 5 draws some conclusions with future direction.

II. PHONETIC SCHEME FOR BANGLA DIGITS

A. Bangla Phonemes

Phonetic inventory of Bangla consists of 14 vowels, including seven nasalized vowels, and 29 consonants. An approximate phonetic scheme in IPA is given in Table I. In Table I (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. Apart from the 29 consonants mentioned in the table, we use another one /ɤ/(/ɐ̯/), which is palatial in place of articulation and approximant in manner.

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].

978-1-4244-6284-1/09/$26.00 ©2009 IEEE

Table I. Bangla phonetic scheme in IPA.[1]

(a) Vowel

	Front	Central	Back
Close	i (ই)		u (উ)
Close-mid	e (এ)		o (ও)
Open-mid	æ (অ্যা)		ɔ (অ)
Open		a (আ)	

(b) Consonants

		Labial	Dental	Alveolar	Apico-Postaveolar	Lamino-Postaveolar	Velar	Glottal
Nasal		m (ম)		n (ন)			ŋ (ং)	
Plosive	Voice less	p (প) pʰ (ফ)	t̪ (ত) t̪ʰ (থ)		t (ট) tʰ (ঠ)	tʃ (চ) tʃʰ (ছ)	k (ক) kʰ (খ)	
	Voiced	b (ব) bʰ (ভ)	d̪ (দ) d̪ʰ (ধ)		ɖ (ড) ɖʰ (ঢ)	dʒ (জ) dʒʰ (ঝ)	g (গ) gʰ (ঘ)	
Fricative				s (স)		ʃ (শ)		h (হ)
Liquid				l (ল) r (র)	ɽ (ড়)			

[1]Some parts are extracted from http://en.wikipedia.org/wiki/Bengali_phonology

Table II. Bangla digit pronunciation.

English Digit	Bangla Digit	Pronunciation (Bangla)	IPA
0	০	শূন্য	/ʃu:nno/
1	১	এক	/æk/
2	২	দুই	/d̪ui/
3	৩	তিন	/t̪in/
4	৪	চার	/tʃar/
5	৫	পাঁচ	/pãtʃ/
6	৬	ছয়	/tʃʰɔ̯/
7	৭	সাত	/sat̪/
8	৮	আট	/at/
9	৯	নয়	/nɔ̯/

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.

B. Bangla Digits

The Bangla script has 10 digits corresponding to the arabic numerals. Table II lists Bangla digits with their written forms and the corresponding ipa. From the table, we can see that '০' (0) and '২' (2) have long /u:/ and short /u/, respectively, though they can be perceived similarly. 'ঁ' (5) has /ã/, which is a nasalized vowel of /a/.

III. BANGLA DIGIT SPEECH CORPUS

At present, a real problem to do experiment on Bangla digit ASR is the lack of proper Bangla digit 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 digit speech corpus, which is described below.

A. Speakers

We have selected 50 male (m01-m50) and 50 female (f01 – f50) - a total of 100 speakers for the corpus. All of the speakers are Bangladeshi residents and native speakers of Bangla. The age of the speakers ranges from 16 to 60 years. We have chosen the speakers from a wide area of Bangladesh: 20 from Dhaka (central region), 10 from Comilla – Noakhali (East region), 10 from Chittagong (South-East region), 20 from Rajshahi (West region), 20 from Dinajpur – Rangpur (North-West region), 10 from Khulna (South-West region), and the rest from Sylhet (North-East region). Though all of them speak in standard Bangla, they are not free from their regional accents.

B. Recording

Recording was done in three quiet rooms located at Dhaka, Rajshahi and Dinajpur. All of the three rooms bore the same type of environment. A laptop was used to record the voices using a head mounted close-talking microphone. All of the speakers spoke ten Bangla digits starting from '০'. We recorded 10 trials of each digit from each speaker: 5 trials in quiet condition and 5 trials in typical Bangladeshi office environment, where ceiling fans were switched on and windows were open, and some low level street or corridor noise could be heard. These two types of trials were recorded in two sessions. A total of 50 utterances were recorded from each speaker in quiet condition (5 trials and 10 digits) and 50 utterances in office environment. The experiment of this paper involves only quiet condition utterances.

GoldWave software was used to record the voices [13]. The speech was sampled at 11025 Hz and quantized to 16 bit coding without any compression. The number of channels was 2, and no filter was used on the recorded voice.

IV. BANGLA DIGIT ASR

Research works on Bangla speech recognition starts from the beginning of this new century. Most of these works involve with vowel and consonant characterization, small vocabulary isolated word recognition, small vocabulary continuous speech recognition, and classification using neural networks. All of the experiments are performed on a small size database, not exceeding 15 speakers [3]-[10].

In this section, we present the first ever Bangla digit ASR on a comparatively larger size database with 100 speakers. The following subsections describes the platform, database, parameters, experiments, and results of Bangla digit ASR.

A. Platform

HMMs are a well-known and widely-used statistical method for characterizing the spectral features of speech frame. HMMs provide a natural and highly reliable way of recognizing speech for a wide range of applications. HTK was used in all the experiments reported here.

B. Database

We divide 100 speakers into training and testing sets. 37 male (m01 – m37) and 37 female (f01 – f37) speakers are used as training set and the rest as testing set. All the five trials for each digit in quiet condition are included in corresponding training and testing sets. Hence we have a total of 2×37×5 instances for each digit in the training set and 2×13×5 in the testing set. In both the training and the testing sets, a balanced representation of speakers from different regions (see Section 3) is maintained.

C. Parameters

The system uses the following parameters: 11.025 kHz sampling rate with 16 bit sample resolution, 25 ms Hamming window with a step size of 10 ms, 13 MFCC features with their first and second order derivatives and 0.97 as the pre-emphasis coefficients.

In addition to these monophones, 'silB' and 'silE' are also used to represent silence at the beginning and at the end of an utterance, respectively. As the size of vocabulary is limited to only 10, we use word models for recognition. Training data set is used to design 10 Bangla digit-HMMs with left-to-right organization. A speaker independent Bangla digit recognition test is then carried out using the testing data set. The test is open as no speaker from the training set is used in the testing set. In the experiment, we vary the number of states from nine to 19, and Gaussian mixture components from one to 12 in each state. We find the optimum result with 13 states excluding nonemitting start and end states, and eight mixture components in each state. We fix these parameters for subsequent experiments.

D. Results And Discussion

We evaluate the system based on digit correct rate (%). Fig. 1 shows digit correct rate (%) of the 10 Bangla digits in the system with 13-state HMMs and varied number of Gaussian mixture components. The correct rate increases with the number of mixture components, however after eight components, it again drops in most of the cases. The following description involves with eight mixture components per state. Fig. 1 demonstrates two categories: one that includes the first six digits with correct rate of over 95% and the other, which includes the last four digits, with correct rate of less than 90%. The highest correct rate (100%) is obtained with the digit '২' (2), while the lowest correct rate (84%) is found with '৮' (8). The low accuracy of the second category is due to the confusion of the system between the members of this category, which is described later.

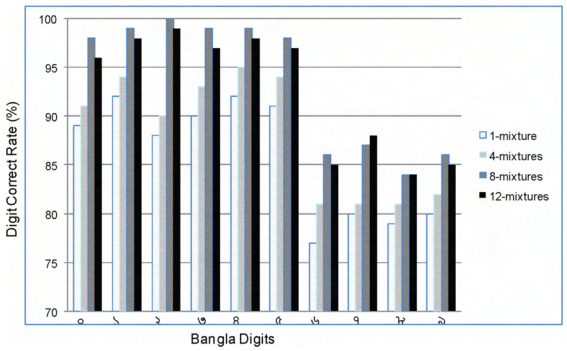

Fig 1. Digit correct rate (%) of Bangla digits in the system.

Table III. 10-digit confusion matrix. The most two confused pairs are: '৬' (6) and '৯' (9); and '৭' (7) and '৮' (8).

Input \ Output	০	১	২	৩	৪	৫	৬	৭	৮	৯
০	98									2
১		99							1	
২			100							
৩				99						1
৪					98			1	1	
৫						98		1	1	
৬							86			14
৭		1						87	12	
৮		2						14	84	
৯			1				13			86

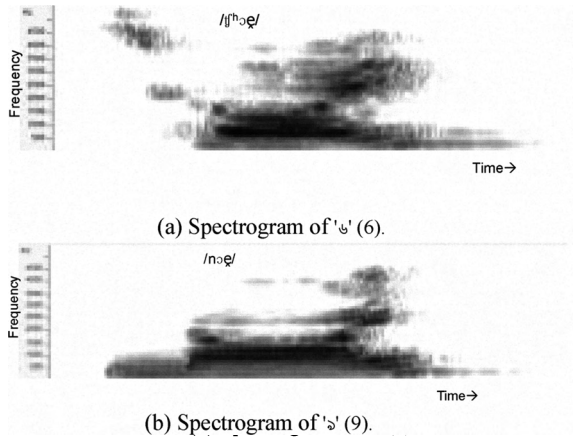

Fig 2. Spectrograms of two confused Bangla digits '৬'(6) and '৯' (9).

Digit '০' (0) has correct rate of 98%, which was reduced to 96% (not shown) while using long /u:/ for its transcription. Optimization of pronunciation dictionary for this digit is thereby justified.

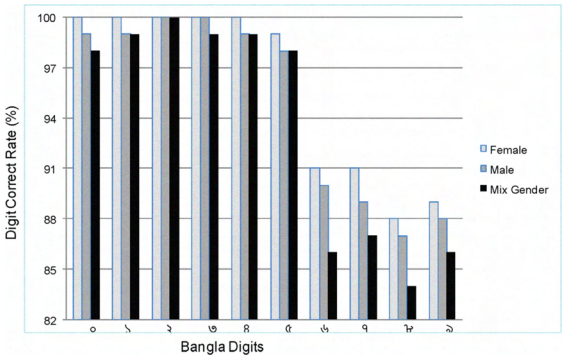

Fig 3. Digit correct rate (%) of Bangla digits in the system (gender dependent). Mix gender result corresponds to the 3rd bar of Fig. 1.

Table III shows confusion matrix generated by the system for all the 10 digits. From the table, we find two most confusing digit pairs: one including digits '৬' (6) and '৯' (9), and the other including digits '৭' (7) and '৮' (8). 14% of '৬' is misclassified as '৯' and 13% of '৯' is misclassified as '৬'. Both of the digits contain long durational phoneme /ɔ/, which contributes most part of the utterances, and different short durational unvoiced parts /ʧ ʰ/ (for '৬') and /n/ (for '৯'). Many of the speakers speak digit '৬' softly without emphasizing the unvoiced parts and this may be a cause for confusion between this pair. Even human listeners sometimes find it difficult to distinguish between softly spoken '৬' and '৯' from long distance. Fig. 2 shows spectrograms of confused pair '৬' and '৯'. From the spectrograms we can see that both digits contain similar phoneme /ɔ/ for most of the part.

Another confusing pair is digits '৭' and '৮'. 12% of '৭' is confused as '৮' and 14% is confused in the reverse direction. Most part of these two utterances has almost similar pattern /at̪/ and /at/. The unvoiced part /s/ in '৭' is very short in duration, and hence sometimes remains undetected. This is one of the reasons for confusion between these two digits.

Fig. 3 gives digit correct rate (%) for gender dependent Bangla digit ASR. 'Mix Gender' in the figure corresponds to the result shown in Fig. 1 with '8-mixtures'. It is obvious than gender dependent ASR has higher correct rate than that with gender independent ASR. In case of female speakers, correct rate for first five digits reaches to 100%, while for male speakers only digits '২' and '৩' have 100% correct rate. From the figure, we can find that, in case of gender dependent, though there is an increase of correct rate for the last four digits, it is still not reaching to the level of the first six digits. Even for some digits within first six, gender dependent cases do not have correct rate of 100%. These observations inspire us to look more insight the last four digits in terms of dialectical difference.

A careful listening of digits uttered by different dialects reveals some clues to improve digit correct rate by the system. For example, many of the natives from

northern part of Bangladesh (Dinajpur – Rangpur) pronounce digit '৪' (4) with nasal voice, which means they pronounce it like /ʧãr/ rather than /ʧar/. In fact, most of the speakers of that region have a tendency to utter with nasalized sound for other digits as well, specially with /a/ sound, for example, /sãt̲/, /ãt/, /æ̃k/. This dialectical difference causes comparatively lower accuracy in ASR for these digits. A preliminary experiment excluding all the utterances from northern region shows a significant improvement of correct rate specially for the digits '৫', '৭', and '৮' (results not shown in figure or table). Therefore, adaptation of different dialects in Bangla digit ASR is our near future goal.

V. CONCLUSION

A Bangla digit ASR system was proposed after developing a medium size Bangla digit speech corpus. The first six digits showed higher accuracy than that of the last four digits. We found two confusing pairs ('৬' (6) and '৯' (9)), and ('৭' (7) and '৮' (8)) in the system. Dialectical difference caused a part of performance degradation. In case of gender dependent experiments, female spoken digits had higher correct rates than those by male spoken digits. In a future work, we will work on adaptation of different dialect in Bangla digit ASR.

REFERENCES

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

[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] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

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

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

[11] S. Young, et al, The HTK Book (for HTK Version. 3.3), Cambridge University Engineering Department, 2005. http:///htk.eng.cam.ac.uk/protdoc/ktkbook.pdf.

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

[13] GoldWave v5.25, available at http://www.goldwave.com/