Developing Bengali Speech Corpus for Phone Recognizer Using Optimum Text Selection Technique

Sandipan Mandal, Biswajit Das, Pabitra Mitra, Anupam Basu
Department of Computer Science
and Engineering
Indian Institute of Technology
Kharagpur, India
mandal.sandipan@gmail.com, biswajit.net@gmail.com, pabitra@gmail.com, anupambas@gmail.com

***Abstract*—Speech corpus plays a key role in construction of automatic speech recognition (ASR), text-to-speech (TTS) synthesis and phone recognition (PR) system. PR system and ASR system are quite similar in functionality. The difference between these two is that for PR system the speech signal is converted to phone[1] text whereas for ASR system the speech signal is converted to word text. Speech corpus for PR system usually consists of a text corpus, recording data corresponding to the text corpus, phonetic representation of the text corpus and a pronunciation dictionary. Selecting optimum text from available text with balanced phone distribution is an important task for developing high quality PR system. In this paper, we describe our text selection technique and discuss the performance of phone recognition system.**

***Keywords*-phoneme; sphinx3; sphinxtrain; MFCC; GMM; HMM**

I. Introduction

Phone recognition is a process to convert speech signal to phonemic textual representation. In spite of Bengali being one of the most popular languages with more than 300 million speakers across the world, there is a lack of research in Bengali speech technology. Most of the Bengali speaking people are from the eastern region of India (West Bengal, Tripura, some parts of Assam and Meghalaya) and Bangladesh[1]. As our experiment is focused on standard Bengali speech, we have not considered dialectical variation due to regional influence.

It is difficult to collect large amount of speech data to get the utterance variation of all phones to build a good PR system. Randomly selected text for speech corpus mostly results in uneven distribution of phone utterance and accuracy of the phone recognizer decreases due to lack of phonetic variation. Optimum text selection process gives an approximate solution to select evenly distributed text with respect to phone frequency. Optimum text selection process also drastically reduces the size of text corpus.

There are several techniques to select optimum text from large text corpus. Set-cover[3] problem is the most popular greedy algorithm[4] for this purpose. In this approach, there is a set of sentences, a parallel set containing list of phones occurring in each sentence of the sentence set and set of phones. The task is to select a subset of sentences from the original sentence set such that the subset contains all phone occurrences at least once[4]. Sentences are selected successively on the basis of largest number of phone count and deleted from the set of sentences. Covered phones are also marked and this selection procedure goes on till all phones are not covered. Some researchers also used threshold-based approach over this greedy algorithm[5]. As number of phones in a speech corpus is limited (47 in Bengali), we need to have the largest number of utterances of less frequent phones in the text corpus to achieve maximum utterance variation. Section II and section III describe the phonetic characteristics of speech corpus and our algorithm for selecting optimum text respectively. In section IV and V we describe overview of phone recognition system and experimental results respectively.

[1]smallest discrete segment of sound in uttered speech

II. Phonetic Characteristic of Corpus

Raw text of our corpus is in unicode. Sentences in the corpus with Bengali unicode fonts are converted to phonemic representation. International Phonetic Alphabet (IPA) corresponding Bengali letter is shown in Table I. During pronunciation of vowel[2], vocal folds vibrates without audible friction and vocal tract configuration remains comparatively open. As the vowel signal is periodic, we can see clear pitch and formant contour. Consonants

Table I
IPA Mapping of Bengali Letters

	Letter	IPA	Letter	IPA	Letter	IPA
Bengali consonant graphemes	ক	/k/	ড	/ɖ/	ম	/m/
	খ	/kʰ/	ঢ	/ɖʰ/	য	/dʒ/
	গ	/g/	ণ	/n/	র	/ɾ/
	ঘ	/gʰ/	ত	/t̪/	ল	/l/
	ঙ	/ŋ/	থ	/t̪ʰ/	শ	/ʃ/ / /s/
	চ	/tʃ/	দ	/d̪/	ষ	/ʃ/
	ছ	/tʃʰ/	ধ	/d̪ʰ/	স	/ʃ/ / /s/
	জ	/dʒ/	ন	/n/	হ	/h/
	ঝ	/dʒʰ/	প	/p/	য়	/e̯/ /-
	ঞ	/n/	ফ	/pʰ/	ড়	/ɽ/
	ট	/ʈ/	ব	/b/	ঢ়	/ɽ/
	ঠ	/ʈʰ/	ভ	/bʰ/		
Vowels	অ	/ɔ/ and /o/	ঈ	/i/	এ	/e/ /æ/
	আ	/a/	ঊ	/u/	ঐ	/oj/
	ই	/i/	ঊ	/u/	ও	/o/
					ঔ	/ow/

978-0-7695-4554-7/11 $26.00 © 2011 IEEE
DOI 10.1109/IALP.2011.16

are pronounced differently. Vocal folds do not vibrate for stop consonants. In case of consonants vocal folds open suddenly and a burst of air comes out.We are unable to get clear pitch and formant frequencies since consonants' signals are not periodic. Pronunciation of word "nak" (নাক) is shown in Figure 1. It clearly describes waveform, spectogram, pitch and formant frequencies for /a/ (vowel), but not for /k/ (consonant)

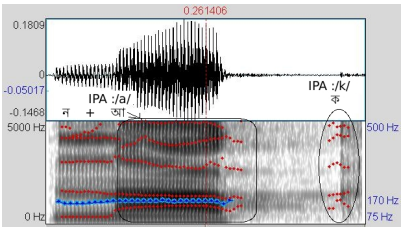

Figure 1. Waveform,Spectogram, Pitch and formant Frequency for /nak/

In our speech corpus, we have selected a total of 47 phonemes which includes 32 consonants, 7 basic vowels and 7 nasalized variation of those vowels. Vowel ae (/ek/ একে) is used frequently. Classification of vowels,consonants are described in Table II and Table III.

Table II
VOWELS OF BENGALI

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

Table III
CONSONANTS OF BENGALI

		Labial	Dental/ Alveolar	Retroflex	Palato-Alveolar	Velar
Nasal		m ,/m/	n, /n/			ŋ, /ng/
Plosive	**voiceless**	p, /p/	t̪, /t/	ʈ,/T/	tʃ,/ch/	k,/ k/
	aspirated	pʰ, /ph/	t̪ʰ,/th/	ʈʰ ,/Th/	tʃʰ, /chh/	kʰ, /kh/
	voiced	b, /b/	d̪,/d/	ɖ,/D/	dʒ,/j/	g, /g/
	aspirated	bʰ,/bh/	d̪ʰ, /dh/	ɖʰ, /Dh/	dʒʰ,/ jh/	gʰ,/ gh/
Fricative		f, /ph/	s,/s/,z,/j/	ʂ,/S/	ʃ, /sh/	
Approximant			l, /l/			
Rhotic			r,/ r/	ɾ,R		
Glottal		h, /h/	Inside two backslash(/ /) our phoneme representation is shown corresponding IPA symbol			

III. PROPOSED ALGORITHM FOR TEXT SELECTION

Initially to design our algorithm for selecting phonetically balanced text for phone recognition system, our target is to maximize less frequent phonemes and minimize more frequent phonemes with minimum text. In our initial text, total no of phonemes is 737552 and standard deviation among 47 phoneme is 19586. All sentences containing phoneme with less frequency (<500) is selected initially. Among all phoneme some of the phoneme frequency in selected sentences is very less. Then selection of the sentences is done one by one with less frequent phoneme with respect to currently selected sentence list. During this selection process, preference is given to shorter sentences so that other phone frequency distribution is kept unchanged as much as possible. People can set their own threshold to other value instead of 500 based on phoneme frequency distribution and corpus size.

This method could be modified after initial sentence selection. Instead of fixing the sentence selection with respect to one of lower value phoneme frequency as threshold, standard deviation (S.D.) of phone frequency can be taken in consideration. Modified algorithm has been described in Algorithm 1.

Algorithm 1: Optimum Text Selection for Bengali Phone Recognition System

Input: All sentences of text corpus, corresponding phone representation of text sentences and phoneme list
Output: Selected sentences with balanced weight of phone frequency
$L \leftarrow$ Empty set of optimally selected sentences;
$M \leftarrow$ Empty set;
$S \leftarrow$ Set of all sentences of text corpus;
$phone\_list \leftarrow$ set of all phoneme;
$len = length(phone\_list)$;
for *i=1 to len* **do**
  $phoneme\_friq[i]$=Number of $phone\_list[i]$ in corpus
end
while $phone\_list[i]$ *is not empty* **do**
  $removed\_phoneme[i]$= Remove a phoneme from $phone\_list[i]$ corresponding minimum val in $phoneme\_friq[i]$;
  foreach *sentence selected in* S *with* $removed\_phoneme[i]$ **do**
    insert *selected* into M;
  end
  for *i=1 to len* **do**
    $phoneme\_friq\_temp[i]$=Number of $phone\_list[i]$ in M;
    Copy M to $Selected\_list[i]$;
    $SD[i]$=Standard Deviation of $phoneme\_friq\_temp[i]$;
  end
end
Copy $Selected\_list[10]$ to L; # In our corpus very less frequent phonemes are till 10th place (<500) in $removed\_phoneme$ list.
for *i=10 to len* **do**
  if $SD[i] < SD[i-1]$ *and all phoneme frequency>0* **then**
    Delete all from L;
    Copy $Selected\_list[i]$ to L;
  end
  else
    L remains unchanged;
  end
end

Table IV
CORPUS SPECIFICATION

Sampling Rate	16 kHz, 16 bits
Format	Wav format
Language,	Bengali
No of Phoneme,	47

This algorithm ensures that the standard deviation of frequency distribution in our optimally selected corpus is minimized. In our corpus, number of least occurring phones are till 10^{th} position in phone frequency list in ascending order. Depending upon the corpus it will vary. Instead of fixing the value to 10^{th} position (in case of our corpus) on the basis of phone frequency distribution, we can only consider set of selected sentences(L in Algorithm 1) with least standard deviation. But in that case, number selected sentences may be very less. So there is a possibility that pronunciation variation of less frequent phone is very less. For this reason, we assumed that all instances of all phones less than 500 phone frequency well be covered.

IV. PHONE RECOGNITION SYSTEM OVERVIEW

SphinxTrain[7] and sphinx3 decoder[7] are used to build the Phone Recognition System.

A. Corpus Creation

To develop speech corpus for Bengali Phone Recognition System text were selected from Anadabazar Patrika, web-based blogs, common conversations and editorials articles. After selecting the optimum text for phone recognition system sentences were recorded in different session and sentences were represented in phonemic format for transcript sentences. Transcript sentences were labeled corresponding to spoken sentences. Corpus specification is given in Table IV.

B. Pronunciation Dictionary

Although alphabet in Bengali pronunciation dictionary contains a total of 29 consonants, 14 vowels and 25 diphthongs[2], it has been observed that pronunciation of our speech corpus can be covered using 47 phonemes. In phone recognition system every phone is treated as a word in pronunciation dictionary.

C. Acoustic Features Computation and Acoustic Model

16-bit 16 KHz wave files are windowed in frames with duration of 25 ms with consecutive frame overlap by 10 ms. Total 39 features are calculated and it consists of 12 Mel Frequency Cepstal Coefficients (MFCC)[6] and 1 energy, 13 first order derivative and 13 second order derivative.

For state probability distribution we use continuous density of Gaussian Mixture distributions. All phonemes are modeled as a sequence of HMM state and likelihoods (emission probability) of a certain frame observation is produced by using traditional Gaussian Mixture Model (GMM)[8].

[2]combination of vowels occurring within the same syllable

D. Trigram Phone Language Model

The probability of any phone in a sequence of phones depends only on the previous N phones in the sequence. We have selected all sentences with phoneme representation from our available corpus to create language model. A trigram phone language model[7] would compute as

$P(wd_1 wd_2 \dots wd_n) =$
$P(wd1)P(wd2|wd1)P(wd3|wd2, wd1)P(wd4|wd3, wd2)...$

V. EXPERIMENTAL RESULTS

Initial text of our corpus contains total 711476 phones. After optimization of text, it contains 143428 phones. Standard deviation of frequency of phonemes for optimized text is 3829, which is very much lower value compared to the value for unoptimized text (19733). The results are shown in Figure 1.

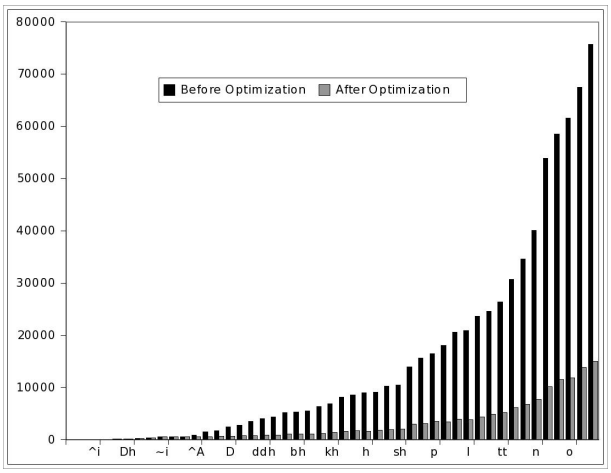

Figure 2. Phoneme Frequency Distribution

Phone recognition with optimally selected text is evaluated with combination different HMM and GMM. The best result is found in 3 state HMM with 16 mixture component for GMM. The comparison is shown in Figure 3.

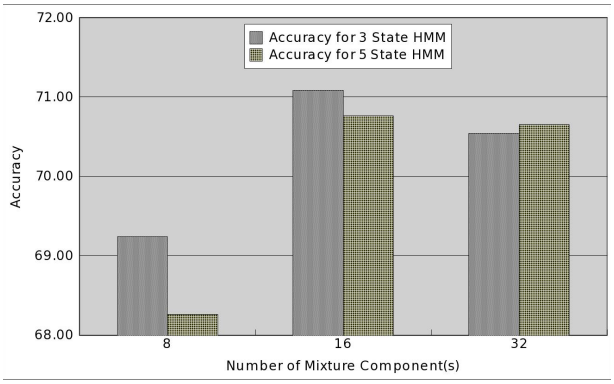

Figure 3. Performance with Different Model

The performance evaluation of phone recognition system has been carried out for greedy text selection method,

Table V
COMPARISON OF SELECTED TEXT BY DIFFERENT APPROACH

Parameter	Initial Text	Greedy Text	Our Text
Total phone	737552	1531	19586
Sentence	13000	10	2000

for randomly selected 3000 sentences and for our optimal text selection method. Random selection method selects some sentences randomly and then selects sentence with uncovered phoneme manually. In greedy text selection method, frequency distribution of some phoneme is very less as number of phoneme in Bengali is 47. As a result probability of positional pronunciation variation is very less. This greedy algorithm may perform significantly good in case of text selection for ASR as triphone and diphone plays the key role instead of phoneme. Same situation arises with thresholdbased algorithm[5]. Table V describes the comparison of number of sentences ,phonemes with respect to initial text, text selected by greedy approach and text selected by our approach.

Results of phone recognition system is shown in Figure 12 using greedy approach,using random approach and using our approach. For this purpose we have chosen 12 speakers (8 males and 4 females) and recorded 92 words which were not in the corpus. Performance of those models has been shown in Figure 4.

As phoneme occurrence in greedy approach is very less and does not have variation in context pronunciation, it gives less accuracy. Although random selection method contains more number of sentences than our approach, number of phonemes with less frequency distribution occur rarely. Thus, our Bengali Phone recognition system gives better performance in our optimum text selection algorithm than the other mentioned methods.

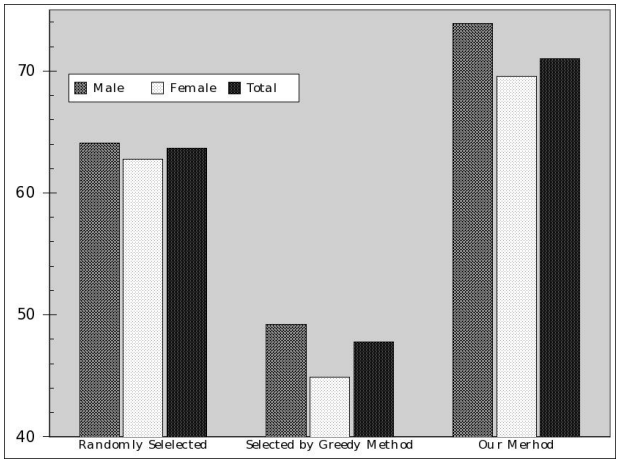

Figure 4. Performance of Phone Recognition System

.

VI. CONCLUSION

Although Bengali is one of the most commonly spoken languages, there is a dearth of research in Bengali ASR and phone recognition. Our text selection algorithm can be used to select optimum text for ASR system using tripone or diphone as a selection parameter instead of phoneme to get better results with less training data. After improving some accuracy of our phone recognition system, it might help to recognize out of vocabulary (OOV) words in speech recognition system.

ACKNOWLEDGMENT

We sincerely acknowledge Communication Empowerment Laboratory, Dept. of Computer Science and Engineering of IIT Kharagpur and Media Lab Asia for providing the funding and setup to carryout our experiment. Thanks to Ratnajit Mukherjee of Communication Empowerment Laboratory for helping us during the evaluation of our experiment.

REFERENCES

[1] *http://en.wikipedia.org/wiki/Bengali_language*.

[2] *http://en.wikipedia.org/wiki/Vowel*.

[3] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms*, 2nd ed. The MIT Press, Cambridge, Massachussetts, 2005.

[4] Jan van Santen and A. Buchsbaum, *Methods for optimal text selection*, Eurospeech97. vol.2,pp. 553556, 1997.

[5] C. Rahul,H. M. Sebsibe, *Rapid methods for Optimal Text Selection*,4em RANLP, 2005.

[6] L. Rabiner, and B. H. Juang, (1993), *Fundamentals of speech recognition*, Prentice-Hall, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[7] CMU Sphinx Group, *http://www.speech.cs.cmu.edu/sphinx/*.

[8] A. Kannan, M. Ostendorf,j. R. Itohlicek, *Maximum Likelihood Clustering of Gaussians for Speech Recognition*, 1994.