BENGALI SPEECH CORPUS FOR CONTINUOUS AUTOMATIC SPEECH RECOGNITION SYSTEM

Biswajit Das

Indian Institute of Technology
Computer Science and Engineering
Kharagpur,India

Sandipan Mandal, Pabitra Mitra

Indian Institute of Technology
Computer Science and Engineering
Kharagpur,India

ABSTRACT

This paper presents Bengali speech corpus development for speaker independent continuous speech recognition. Speech corpora is the backbone of automatic speech recognition (ASR) system. Speech corpus can be classified into several class. It may be language dependent or age dependent. We have developed speech corpus for two age groups. Younger group belongs to 20 to 40 years of age whereas older group is distributed into 60 to 80 years. We have created phone and triphone labeled speech corpora. Initially, speech samples are aligned with statistical modeling technique. Statistically labeled files are then pruned by manual correction. Hidden Markov Model Toolkit (HTK) has been used for aligning the speech data. We have observed phoneme recognition and continuous word recognition performance to check speech corpus quality.

***Index Terms*—** HTK, SPHINX, Speech labeling, Bengali speech corpus, Speech recognition.

1. INTRODUCTION

There are about 250 million people speaks in Bengali. Speech Recognition involves the understanding of the naturally spoken speech by the native users, its semantics as well as meaning particular to a context. The contextual meaning helps building a better language model for automatic speech recognizers. Different statistical models are used for speech recognition by the researchers.

For having better performance of the speech recognizers, it is inevitable to have speech corpus of that particular language. The first initiative was taken by MIT to create TIMIT speech corpus [1]. Speech corpus has been developed in many other foreign languages [2, 3]. Creation of a speech corpus is a mandatory task for building and testing any speech recognition system. The quality of the recognizer depends on the quality of the speech corpus also. There can be a number of reasons that can drastically change the performance of a speech recognition system. The reasons are like session variability, inter-speaker variability and instrument variability. The speakers can utter same words in different manners during different session and different emotional state. This can affect the baseline performance of a speech recognition system very much. Moreover another important issue is dialects of the language. It is very much important to benchmark the system with a standard database, which covers a good range of variability. A wide variety of languages belonging to different linguistics group are spoken in India. In addition, different dialects are spoken in the same language depending on the region to which speaker's belong.

Different applications for automatic speech recognition use speech as input for the system but the accuracy reduces to a great extent if there exists speech variability between test samples and train samples. Moreover, in case of speech recognition system to be built in Bengali, speech database is the most essential part for training of the system and for benchmarking the system also. The project ”SHRUTI” and ”ERA” deals with automatic speech recognition system. This motivated us to design and create a speech corpus in standard colloquial Bengali language for older as well as younger people.

We have used HTK and CMU-SPHINX speech recognition toolkit to train and test speech recognition system. Monophone model is created with HTK to test phoneme recognition accuracy and triphone model is developed with SPHINX to test word recognition. Speech recognition performance affected much due to inter-speaker variability. Speaker variability can be reduced with speaker normalization techniques. Speech features can be normalized in cepstral domain by means of cepstral mean and variance normalization (CMN). Speaker adaptation method like Maximum likelihood Linear Regression (MLLR) [4] is applied at the time of acoustic modeling.

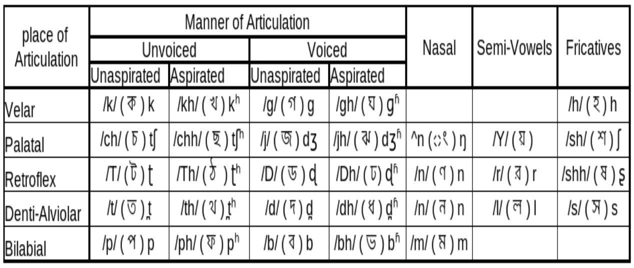
2. BENGALI PHONETIC CHARACTERISTICS

There are regional variation in Bengali language called dialect. Rarh, Banga, Kamarupa and Varendra are four type of well known dialects in Bengali. Rarh is the standard colloquial language in south-western region. Even in Rarh dialect, there are sub-regional variation. Style of pronunciation and

978-1-4577-0931-9/11/$26.00 ©2011 IEEE

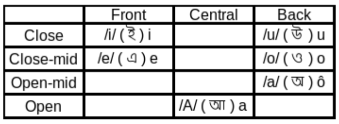
accent are different for different speakers. Same word is pronounced differently by different speakers.

A wide variety of languages belonging to different linguistics group are spoken in India. In addition, different dialects are spoken in the same language depending on the region to which speaker's belong. Considering all these issues, we have selected 47 phonemes for Bengali. Bengali phonemes are classified with some unique phonetic properties [5]. The properties are described as place of excitation, manner of excitation and type of excitation. On the basis of articulatory system phonemes are designed. Phonemes can be divided as consonant, vowel and semi-vowel. Phone classification depends on the following question. Duration of the phones are short or long, position of articulation, voiced or unvoiced. Different manner of articulation like nasal, fricative, affricative or stop. Point of articulation also differentiate the phonemes like labial, dental, bilabial, velar, alveolar or alveolar-dental. We have represented the phoneme in / /, Bengali alphabet in () and International Phonetic Alphabet (IPA) for consonants and vowels respectively in the Figure 1 and 2.



place of Articulation	Manner of Articulation				Nasal	Semi-Vowels	Fricatives
	Unvoiced		Voiced				
	Unaspirated	Aspirated	Unaspirated	Aspirated			
Velar	/k/ (ক) k	/kh/ (খ) kʰ	/g/ (গ) g	/gh/ (ঘ) gʱ			/h/ (হ) h
Palatal	/ch/ (চ) tʃ	/chh/ (ছ) tʃʰ	/j/ (জ) dʒ	/jh/ (ঝ) dʒʱ	^n (◌ং) ŋ	/Y/ (য)	/sh/ (শ) ʃ
Retroflex	/T/ (ট) ʈ	/Th/ (ঠ) ʈʰ	/D/ (ড) ɖ	/Dh/ (ঢ) ɖʱ	/n/ (ণ) n	/r/ (র) r	/shh/ (ষ) ʂ
Denti-Alviolar	/t/ (ত) t̪	/th/ (থ) t̪ʰ	/d/ (দ) d̪	/dh/ (ধ) d̪ʱ	/n/ (ন) n	/l/ (ল) l	/s/ (স) s
Bilabial	/p/ (প) p	/ph/ (ফ) pʰ	/b/ (ব) b	/bh/ (ভ) bʱ	/m/ (ম) m		

Fig. 1. Bengali consonant phoneme characteristics



	Front	Central	Back
Close	/i/ (ই) i		/u/ (উ) u
Close-mid	/e/ (এ) e		/o/ (ও) o
Open-mid			/a/ (অ) ô
Open		/A/ (আ) a	

Fig. 2. Bengali vowel phoneme characteristics

3. CORPUS

There can be a number of reasons that can drastically change the performance of a speech recognition system. The reasons are like session variability, intra-speaker and inter-speaker variability and instrument variability. The speakers can utter same words in different manners during different session and different emotional state. This can affect the baseline performance of a speech recognition system very much. Moreover another important issue is dialects of the language. It is very much important to benchmark the system with a standard database, which covers a good range of variability. We have used optimal text selection technique for building up phonetically balanced text corpora. Text corpora consist of 7500 unique sentences, 19640 unique words. Sentences are recorded by 70 male speaker and 40 female speaker. Sentences are collected from various domain of Anandabazar text corpora. Each sentences are recorded with sample frequency 16000 Hz and mono channel. Speech signals are encoded with 16 bit encoding. Sony FV-220 microphone and Emu speech tool has been used for speech recording. Dictionary is a phonetic representation of words. Dictionary has been created with grapheme to phoneme conversion procedure. After this, dictionary is corrected manually. We have developed 26 hours continuous read speech corpus.

We have collected data from two types of age group in room environment. First group is belongs to 20 to 40 years of age. Second group is chosen from 60 to 80 years. There are 30 young male and 40 older male speakers in the men group and 20 young female and 20 older female speakers in the corpus. There are few Bengali speech corpora for younger people but in different dialect. It is first time to develop continuous read speech corpus for older people. Speech data collection from older people is a challenging task. Most of the older people suffers due to teeth loss and low vision. It creates problem at the time of pronunciation. In general, speech quality degrades with aging. Physiological changes in articulatory system affect the voice source features. Teeth loss, tongue muscle strength loss and other physical changes create problems to pronounce specific sound. Voice source features like Fundamental frequency (F0), formant frequencies, Jitter, Shimmer, Voice onset time (VOT) and Harmonic-to-Noise ratio (HNR) are changes significantly with aging [6].

Test data set is very much essential for evaluating ASR performance as well as corpus characteristics. We have recorded different 20 sentences which are not used in training corpus. Speakers are also different from training data set to proof speaker independency. There are 20 speakers(10 younger and 10 older) in the test corpus. Those 20 sentences covers all phonemes which are used in training data set.

Speech corpus has different parts. It consist of speech sample, dictionary, phoneme list and transcription file. Dictionary presents phonetic representation of each words. We have created the pronunciation dictionary manually because same word can be pronounced differently.

Word	Phoneme							
abasthAYa	a	b	o	s	th	A	Y	
abasthita	a	b	o	s	th	i	t	o
Abedana	A	b	e	d	a	n		
abhAba	a	bh	A	b				

Table 1. Bengali pronunciation dictionary

Phoneme distribution in the text corpus is described in Figure 3. We have seen that some phoneme are used rarely in

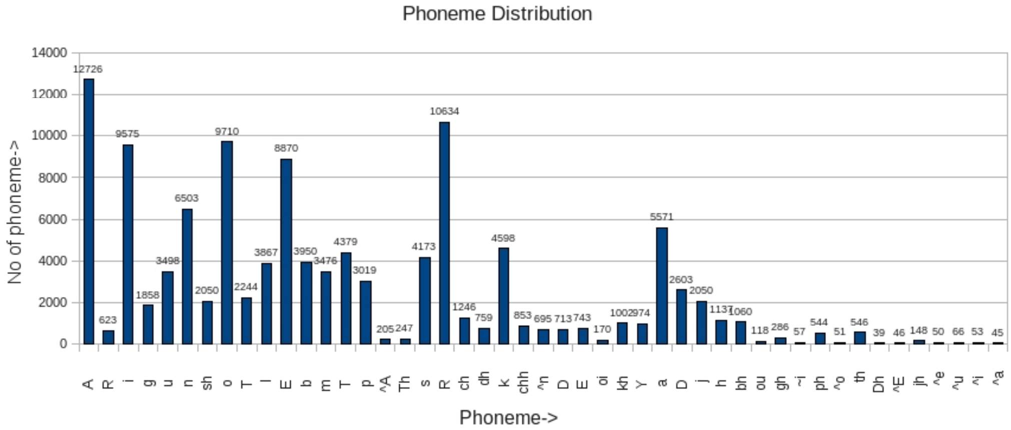


Fig. 3. Phoneme distribution in the text corpus

Bengali language. We have separated nasal vowels and normal vowels in this study. Though they are distinct by nature from each other, Nasal vowels are less in number.

4. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition (ASR) [7] is the process of converting speech into text. There are three main segment of ASR. First step is to extract the feature from each speech signal. In the next step, Acoustic modeling has been done for each phoneme with the feature file. Last segment of the ASR system is language model. It captures the linguistic information from the text.

4.1. Feature extraction

Speech is a non-stationary signal. Speech may be assumed stationary for a short time period(20-30 msec) called frame. We are used to extract information from this frames by means of different feature extraction procedure. Linear Prediction Cepstral Coefficient (LPCC) is process of estimating current speech sample by approximating the past speech sample. Mel Frequency Cepstral coefficient (MFCC) [8] is the another most popular feature extraction procedure in cepstral domain. In this process, Frequency components are mapped with Mel scale. There is another feature extraction procedure called Perpetual Linear Prediction (PLP). Energy spectrum of each frames are mapped with Bark scale. In this study, We have used MFCC feature.

4.1.1. MFCC feature

We converts each frame to frequency domain with Fast Fourier Transform (FFT). After having spectral feature from speech signal, frequencies are converted to Mel frequencies. Number of triangular band-pass filter is placed in mel-frequency scale within a band. MFCC features are computed with discrete cosine transform (DCT) using the filter output amplitude. Each frame is parametrized to feature vector. It consist of base MFCC features plus their first and second order time derivatives. It has the benefit that it is capable of capturing the phonetically important characteristics of speech. We have used two types of MFCC feature set.

First set consist of 26 dimensional feature vector. We have selected 13 base feature with it's first order derivative for it. Second set consist of 39 dimensional feature vector for each frame. It is a combination of 13 base feature with it's first and second order derivatives.

4.2. Acoustic model

We have used two types of toolkit for acoustic modeling. HTK has been used for phoneme recognition without language model. We have used only the model parameter for phoneme recognition. Same corpus has been used to model triphone with CMU SPHINX toolkit. In this case, trigram language model has been used.

Training procedure starts with estimating the global mean and variance. After flat initialization of HMM parameter, Re-estimation of HMM parameters for each phoneme are computed with Baum-Welch re-estimation. In this study, we have created two types of context related model. At first, Context independent monophone HMM model has been done. We have selected 5 state left-to-right HMM model. Each state is represented with 8 Gaussian components. In case of continuous speech production mechanism, current phone is influenced by it's past and future phoneme. This is called co-articulation effect. We have used triphone as a basic unit to model co-articulation effect. Context dependent unit increases the model set. As the training data set is not sufficient to cover all triphones, Decision tree based state tying procedure is applied to reduce the total number of parameter. After completion of monophone model, training speech samples are aligned with respect to phone. It provides phone labeled training speech corpus. Phone labeled speech samples are manually corrected to get final phone labeled corpus. In the same way, we have got triphone labeled speech corpus.

We have used Maximum Likelihood Estimation (MLE) based tied triphone acoustic model for continuous speech recognition. Feature space transformation is done with Linear Discriminative Analysis (LDA). It improves separability of acoustic classes in the feature space. We have reduced the dimensionality of features from 39 to 29 with LDA. These transformed features are used in MLLR based speaker adaptive acoustic modeling.

4.3. Language model

Language model captures the contextual information in the training text corpus. We have not used any language model for phoneme recognition in HTK but only a phone network has been used. In case of continuous speech recognition, trigram language model has been implemented. It is a probability score of a word given it's previous two words. It improves the ASR performance.

5. RESULTS

ASR performance reflects the quality of speech corpus. We shall first analyze the phoneme recognition performance with different test feature vector as well as different model parameter. We have discussed two types of cepstral feature vector in previous section. It is clear from the recognition performance that 39 dimensional feature performs better than 26 dimensional feature. We have created separate acoustic model with training speech data from aged and young population. We have tested test data of young population with acoustic model parameter of young. Test data of aged group is tested with model parameter of young as well as aged. Test data of young people gives good performance with young model parameter but performance degrades with test data of old people. Phone recognition of older people improves with model parameter of old people. Different test set are described below.

- YM_YT_39: Model of young people with 39 dimensional feature and test data from young.
- YM_OT_39: Model of young people with 39 dimensional feature and test data from old.
- OM_OT_39: Model of aging people with 39 dimensional feature and test data from old.
- YM_YT_26: Model of young people with 26 dimensional feature and test data from young.
- YM_OT_26: Model of young people with 26 dimensional feature and test data from old.
- OM_OT_26: Model of aging people with 26 dimensional feature and test data from old.

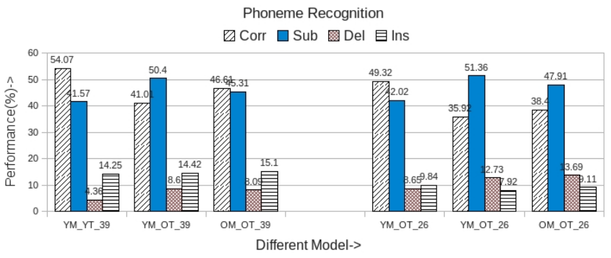


Fig. 4. Phoneme recognition performance

We have tested word recognition with same test data set. We have shown the correctly(%) recognized words with continuous test data. We have used only 39 dimensional feature for continuous speech recognition. Recognition rate is better for young test data. Recognition accuracy degrades much for aging test data. Specific model for aging people improves the performance. We have used LDA and MLLR based speaker normalization.

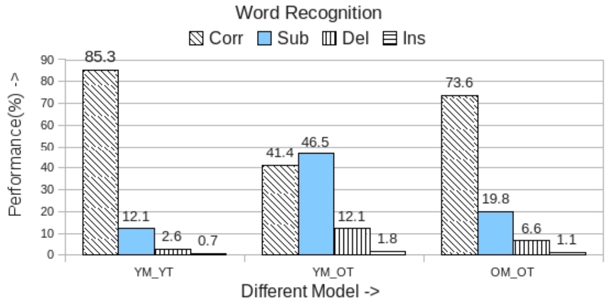


Fig. 5. Word recognition performance

6. CONCLUSION

In this paper, we have build a continuous read speech corpus of young and old people. Most of the speech corpus has been developed in any language are for young people. We have developed speech corpus of standard colloquial Bengali language which is mostly spoken in West Bengal, India. ASR performance degrades due to speaker variability, environment noise and transmission channel noise. Speaker variability increases among young and old population.

Voice source feature like F_0, formant frequencies, jitter, shimmer, HNR and VOT are changes more or less with aging. These changes affect ASR performance more and less. As Speech features differ more among young and old, specific model is required for each group. It will improve ASR performance.

Phoneme recognition with monophone model and phone network gives good accuracy. We have used two types of MFCC features for modeling each phone. 26 dimensional feature is constructed with one cepstral energy and 12 MFCC and 13 delta feature. 39 dimensional feature set is consist of 13 base MFCC feature and it's delta and delta-delta feature.

We have developed also a continuous word recognition system with CMU SPHINX. ASR performance is comparable with ASR systems of other languages. We have used MLLR speaker normalization technique to improve the ASR performance. 39 dimensional feature are converted to 29 dimensional feature with LDA. Test data of younger people gives better result than older.

This speech corpus can be used for age detection work. As we have covered long range of age, Age detection model can be done with this corpus.

7. ACKNOWLEDGMENTS

This work was supported by Technology Intervention for Elderly, Department of Science and Technology (DST), Government of India and Media Lab Asia. I am glad to thanks Prof. Anupam Basu and Suparnakanti Das for their inspiration. Finally, I wish to thanks all individuals of Communication Empowerment Laboratory, Indian Institute of Technology, who have participated in this study.

8. REFERENCES

[1] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351 – 356, 1990.

[2] R. Carre, R. Descout, M. Eskenazi, J. Mariani, and M. Rossi, "The French language database: defining, planning, and recording a large database," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.* IEEE, 1984, vol. 9, pp. 324–327.

[3] Takeda K. Sagisaka Y. Katagiri S. Morikawa S. Watanabe T. Kuwabara, H., "Construction of a large-scale Japanese speech database and its management system ," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'89.* IEEE, 1989, vol. 1, pp. 560 – 563.

[4] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[5] Suniti Kumar Chatterji, "Bengali phonetics," *Bulletin of the School of Oriental Studies, University of London*, vol. 2, no. 1, pp. 1–25, 1921.

[6] Ravichander Vipperla, Steve Renals, and Joe Frankel, "Ageing voices: The effect of changes in voice parameters on asr performance," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010.

[7] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257 –286, feb 1989.

[8] J. S. Bridle and M. D. Brown, "An experimental automatic word recognition system," *JSRU Report No. 1003*, 1974.