Ethnographically Oriented Repository of Assamese Telephonic Speech

Mridusmita Sharma[1,*], *Kandarpa Kumar Sarma*[2], and Nikos E. Mastorakis[1]

[1]Department of Electronics and Communication Engineering, Gauhati University, Guwahati-781014, Assam, India

[2]Department of Electronics and Communication Engineering, Gauhati University, Guwahati-781014, Assam, India

[3]Technical University of Sofia, Sofia, Kliment Ohridski 8, Bulgaria

Abstract. Recording of the speech samples is the first step in speech recognition and related tasks. For English, there are a bunch of readily available data sets. But standard data sets with regional dialect and mood variations are not available and the need to create our own data set for our experimental works has been faced. We have considered Assamese language for our case study and since it is less computationally aware, there is a need to develop the speech corpus having dialect and mood variations. Also, the development of corpus is an ongoing process and the initial task is reported in this paper.

1 INTRODUCTION

The primary objective of designing an efficient Automatic Speech Recognition (ASR) system lies in its superior performance over the already existing models [1]. The performance evaluation of an ASR system greatly depends upon the speech database which is the core part of the system. Successful deployment of an ASR system in real time requires a versatile and relevant database [2]. There are a bunch of readily available datasets for major languages like English. But standard data sets with regional dialect and mood variations are not available and the need to create our own data set for the experimental work has been faced.

Recognition of speech in the regional dialects permits the removal of digital divide among the computer illiterate people and the people with good computer knowledge. Dialect recognition hence permits a huge section of the population to use the benefits of technology [3]. Research in the field of speaker recognition has also evolved at par with speech recognition and speech synthesis because of the similar characteristics and challenges associated with it. Research and development in the field of speaker recognition dates back to the last century and this area is still an active topic for research. The application of speaker recognition technology has been continually growing in various fields of application such as forensic applications, dynamic signatures, gait, keystroke recognition, data encryption purpose, user validation in contact centers [4]. Also, the use of telephonic speech is getting more involvement in many new applications in spoken language processing. However, there are significant challenges that need solutions while designing system for real time telephone speech recognition with better accuracy with regional linguistic orientation. One such challenge is the design of a good speech corpus which has regional dialectal and mood variations. A balanced corpus is the basic requirement for designing a mood or dialect or speaker recognition system.

For designing a speech database, the following considerations may be taken into account-

1. Depending upon the scope and the problems identified, the corpora can be designed accordingly. For eg, dialect recognition, emotion recognition, speaker recognition, etc.
2. The dataset can have variety of contents depending upon the requirements, like short sentences to long sentences, telephonic recordings, normal recordings. Data may be collected in different sessions to incorporate intersession variability.
3. The total number of samples or speakers should be enough to validate the experiment. Also, for ensuring the accuracy of proper pronunciation, the utterances of each samples should be spoken more than once.
4. The utterances may be recorded in different environments like recording studio, noise free closed room, office environment, outdoor environment, etc.

As already mentioned, Assamese is not a computationally popular language and to build an ASR system having regional and dialectal variations permits

* Corresponding author: mriduzb@gmail.com

to remove the digital divide among the computer illiterate people and the people with good computer knowledge. In this paper, an Assamese database having dialectal, emotional, speaker and telephonic variations is described that is used for our work. We have used this dataset for emotion, dialect and speaker recognition purpose.

The rest of the paper is organized in the following sections. Section II gives a description of the Assamese language and its dialects. Section III describes the collection of the Assamese speech corpus. Section IV shows some of the experimental results of various problems obtained by using the corpus. Section V concludes the paper.

2 The Assamese Language

Assamese is the dominant language of Assam which is an Indo-Aryan language belonging to the Indo-European language family. The Assamese language developed from the Apabhramsa dialects developed from Magadhi Prakrit of the eastern group of Sanskritic languages. It is spoken by over 15 million native speakers and is also spoken in Arunachal Pradesh and other north-eastern states of India. Assamese is the official language of the state of Assam and is one of the 23 official languages recognized by the Republic of India. The phonemic inventory of the Assamese language consist of 8 vowels, 10 diphthongs and 23 consonants. In earlier days, Banikanta Kakati has divided the Assamese language into two major groups or dialects: Eastern group and Western group with respect to the two ghats of the Brahmaputra valley [5] [6]. However, because of certain external influences, recent studies have shown that there are four major dialects of Assamese language:

- Eastern Assamese dialect: spoken in the districts of Tinsukia, Dibrugarh, Lakhimpur, Dhemaji, Sibsagar, Jorhat, Golaghat and Sonitpur.
- Central Assamese dialect: spoken primarily in Nagoan and Morigaon districts and in some parts of Sonitpur and Jorhat districts.
- Kamrupi Assamese dialect: spoken in the districts of Kamrup, Nalbari, Barpeta, Darrang, Kokrajhar and Bongaigoan, and
- Goalpariya Assamese dialect: spoken primarily in the Dhubri and Golapara districtsa and in certain areas of Kokrajhar and Bongaigoan districts.

3 Creation of Database

As mentioned earlier, since there is no standard built-in database available in Assamese language and since we have considered Assamese language for our case study, we have created the database as per our requirement for the experimental work.

To be able to work on the problems identified and test the proposed recognition models and feature extraction techniques, we have recorded the Assamese telephonic speech having mood and dialect variations.

We have prepared a list of sentences which comprises of two, three, four, five and six word Assamese sentences. The sentences are chosen in such a way so as to get a reasonable diverse set of data. These sentences are then recorded by the native speakers of the region. For recording the speech having dialectal variations, we have considered the four major classes of the language namely Goalpariya, Kamrupi, Central and Eastern dialects. The speakers have been asked to infuse the mood variations while recording. The moods considered in our case is angry, loud and normal. We have considered three speakers for each dialect and one sentence have been uttered 5 times. The recorded audio have been properly tagged according to the word-length, emotion, dialect, sentence identity, etc.

For a speaker recognition task using i-vectors as features, a long Assamese sentence consisting of 17 words has been prepared having considerable variations. We have considered fifteen speakers of both male and female genders for recording the samples. For incorporating the speaker and the channel variability, the recorded samples consists of four channels which comprises of telephonic recordings and high quality normal recordings.

After completing the recording sessions, the audio files of each utterances are labeled and separated. Each audio file contains silence part in the file. The silence part present in the audio is the silence captured during recording. Before proceeding towards feature extraction and other stages of the system model, preprocessing of the recorded samples have been carried out. The recorded audio files were saved in .WAV format which is required for further processing.

The work has been carried out in a workstation with Intel Xeon processor @ 3.10 GHz with 16.0 GB RAM. The labeling of the samples has been carried out using the speech analysis software PRAAT [7]. The silence region present in the speech samples have been removed manually using the PRAAT software and also automatically by designing an algorithm. Table 1 gives an overview of the database collected for our experimental work.

Table 1. Overview of the Assamese speech database.

Sl No.	Classification Types	Remarks
1	Mood and Dialect Recognition	4 Dialects (Kamrupi, Goalpariya, Central and Eastern), 3 moods (angry, loud and normal), 2, 3, 4, 5, 6 word Assamese sentences with five utterances per speaker.
2	Mood and Dialect Recognition	Time shifted examples (1 sec, 2 sec, 3 sec and 0.8 sec delays) of all the samples of Sl. No. 1
3	Speaker Recognition	15 speakers each of male and female volunteers with 4 dialects variations with 17 words sentences. Each utterance is recorded using high quality digital USB microphone, normal quality recording using laptop mic and recording using mobile phone.

3.1. Speaker Selection

During the recording of the speech samples, the following considerations were made while selecting the male and female speakers.

- Native Speakers: All the speakers considered were the native of the state of Assam and the belong to the respective dialect groups.
- Speaker's age: The age of the speaker's ranges from 22 to 35 years.
- Educational qualifications: The speakers were mainly post graduate students of Gauhati University.

3.2. Recording Specifications

We have recorded the speech samples in a noise free environment in the Department of electronics and Communication Technology, Gauhati University.
We have used high quality digital USB headset, normal quality recording using laptop mic and recording using mobile phone. The distance of the mic and the speaker's mouth were kept to be 10 cm approximately and the mouth of the speaker were at per level.

The database recorded using mobile phone were sampled at 8000 Hz. Table 2 gives the specifications that were considered while recording the speech samples.

Table 2. Specifications of Recording.

Recording Software	PRAAT and Mobile phone (android OS)
Sampling frequency	8000 Hz for mobile recording; 16000 Hz for normal recording
Channel	mono

Figures 1 and 2 shows the recording setup used for the recording the speech samples using headset and mobile handset.

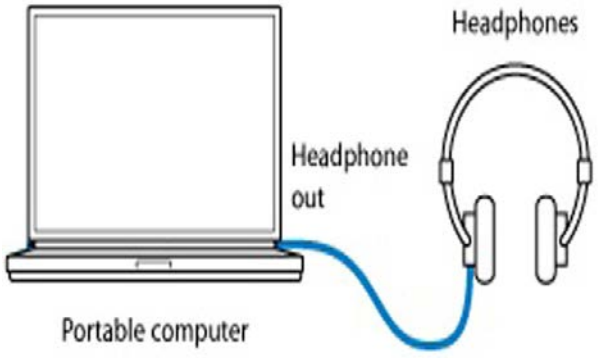

Fig. 1. Setup for normal recording using high quality headphone.

Fig. 2. Setup for normal recording using high quality mobile handset.

4 Experimental Works Using the Database Prepared

For Speaker recognition, we have used the MLFNN and i-vector combination for classification using the recorded

database mentioned in this paper. The results obtained is summarized in Table 3.

Table 3. Speaker Recognition using the recorded Assamese database.

System model	Work	Accuracy (%)	Computational time (secs)	Advantage
i-vector + MLFNN	Our work	87.5 to 98.8	5.56 to 29.28	The proposed system has an improved accuracy by 4.3 %
i-vetor + Gaussian classifiers	[8]	67.1 to 94.5	1 to 3	

For the emotion recognition task, we have used three moods, angry, loud and normal. The emotion recognition model comprises of composite i-vector features and RNN and TDNN classifiers. The results obtained from the experiment is summarized in Table 4.

Table 4. Emotion Recognition using the recorded Assamese dataset.

System model	Work	Accuracy (%)	Computational time (sec)	Advantage
RNN + i-vector Composite features	Our work	86.55	85.6	Comparing the results with [9], we observed that our proposed emotion recognition system has an improved recognition accuracy of .53 %. Comparing the results with [10], we observe that our proposed system outperforms their system by an increase in the accuracy rate by 37.65 %.
RNN+ MFCC-Delta features		83.65	66.95	
TDNN+ i-vector Composite features		79.9	6.15	
TDNN+ MFCC-Delta features		73.85	66.25	
GMM+ LPCC+ i-vector	[9]	86.02	1.81	
RNN+i-vector	[10]	48.9	65.2	

As already mentioned, there are four major dialects of Assamese language and for the purpose of dialect recognition, we have accordingly recorded the speech samples from the native speakers of the districts with respect to the dialects. The results obtained by using our dataset for dialect recognition is compared with some of the already existing models and the summary of the results is tabulated in Table 5.

Table 5. Dialect Recognition using the recorded Assamese dataset.

System model	Work	Accuracy (%)	Computational time (sec)	Advantage
RNN+ Fourier parameters	Our work	95	27.44	For dialect recognition, our proposed system is better than the existing systems by an accuracy of 14 % approx.
TDNN+ Fourier parameters		92.5	105.08	
SVM + Fourier parameters		87.5	1.39	
ANN + Spectral features	[11]	78	69.5	
SVM + Spectral features	[11]	81	1.3	

From the experimental results summarized in Tables 3, 4 and 5, it is clear that the Assamese dataset developed by us works well on the proposed system models for speaker, emotion and dialect recognition.

5 Conclusion

In this paper, we have reported the development of an Assamese dataset having dialectal, mood and speaker variations. The database is recorded using high quality headsets for normal recording and mobile handsets for telephonic recordings. The dataset used in various dialect, emotion and speaker recognition tasks provided satisfactory recognition accuracy and from the experiment, the recognition models were found to be advantageous than some of the existing models. However, the development of a corpus is an ongoing process and it is still in its initial stage. Once a fully

developed dataset is created, it will significantly contribute towards the Assamese regional speech processing tasks.

Acknowledgement

Authors are very grateful to MEITY (Visvesvaraya PhD Fellowship) for assisting financial support to do the research work.

References

[1] M. A. Anusuya and S. K. Katti, *Speech Recognition by Machine, a Review*, International Journal of Computer Science and Information Security, **6,** 3, pp. 181-205, (2009).

[2] F. Ehsani and E. Knodt, *Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm*, Language Learning & Technology, **2**, 1, pp. 54-73, (1998).

[3] M. Sarma and K. K. Sarma, *Phoneme-Based Speech Segmentation Using Hybrid Soft Computing Framework*, Studies in Computational Intelligence, **550**, India, (2014).

[4] C. Kurain, *A Survey on Speech Recognition in Indian Languages*, International Journal of Computer Science and Information Technologies, **5**, 5, pp. 6169-6175, (2014).

[5] M. Sharma and K. K. Sarma, *Dialectal Assamese Vowel Speech Detection using Acoustic Phonetic Features, KNN and RNN*, in proceedings of IEEE 2nd International Conference on Signal Processing and Integrated Networks (SPIN), pp. 674-678, IEEE, (2015).

[6] G. C. Goswami, Structure of Assamese, 1st Ed., Department of Publication, Gauhati University, (1982).

[7] P. Boersma and D. Weenink, *Introductory Tutorial to Praat*, (2011).

[8] D. A. Reynolds, *Speaker Identification and Verification using Gaussian Mixture Speaker Models*, Speech Communication, **17**, 1-2, pp. 91-108, (1995).

[9] L. Mackova, A. Cizmar and J. Juhar, *Emotion Recognition in i-vector Space*, in 26th International Conference Radioelektronika (RADIOELEKTRONIKA), pp. 372-375, IEEE, (2016).

[10] T. Zhang and J. Wu, *Speech Emotion Recognition with i-vector Feature and RNN Model*, in 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), pp. 524-528, IEEE, (2015).

[11] K. S. Rao and S. G. Koolagudi, *Identification of Hindi Dialects and Emotions using Spectral and Prosodic Features of Speech*, IJSCI: International Journal of Systemics, Cybernetics and Informatics, **9**, 4, pp. 24-33, (2011).