Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment

Utpal Bhattacharjee, Kshirod Sarmah

Abstract*— Automatic Speaker Verification (ASV) refers to the task of verifying the claimed identity of a speaker based on speech data. The decision made by a Speaker Verification system is basically a binary decision returns either "Yes" or "No" based on the credibility of the claim, determined by some scoring techniques. The output of an automatic speaker verification system is highly dependent on database used for training and testing the system. The results obtained by the speaker verification system are meaningless if recording specifications and environment for training and testing data are not known. This paper describes methodology and experimental setup used for the development of a speech corpus for the evaluation of text-independent speaker verification system in multilingual environment. Four major languages of Arunachal Pradesh (a North-Eastern frontier state of India, boarding with China) Nyishi, Adi, Galo and Apatani along with English and Hindi have been considered for the developing of the speech corpus. Each speaker has been recoded for three languages – English, Hindi and a local language which must be the mother tongue of the informant. A basic characteristic of this corpus is the presence of both native and non-native speaker. English and Hindi languages have been considered as non-native languages for the speaker. Though the corpus is basically developed for the speaker and language recognition research, it can also be used for various studies including the influence of non-nativeness on speaker and language recognition and accent recognition.

Keywords- Speaker Verification, Speech corpus, Multilingual, Non-native.

I. INTRODUCTION

The general area of Automatic Speaker Recognition encompasses two fundamental tasks – Automatic Speaker Identification (ASI) and Automatic Speaker Verification (ASV). Automatic Speaker Identification is the task of determining who is talking from a set of known voices or speech with the assumption that the unknown voice must come from a fixed set of known speakers. Thus, the task is often referred to as Closed set Identification. Automatic Speaker Verification is the task of determining whether a person is who he/she claims to be (a Yes/No decision). Since it is generally assumed that imposter are not known to the system, this is also referred to as an Open set Verification task. The modern day speaker recognition systems consists of six key components[1]. The components are: (i) Filtering and Analog-to-Digital (A/D) Conversion (ii) Silence Removal (iii) Front-end Processing (iv) Pattern Matching (v) Decision logic and (vi) Enrollment. The filtering and A/D section is responsible for capturing speech from the real world. The silence is then removed from the speech and converted into a series of highly representative short-time spectral features that highlight the speaker specific properties present in the speech. Using these features the pattern matching section relates them to stored models and calculates a distortion/probability for each model. Using the result of the pattern-matching section the system makes a decision on the validity of the speaker's claim or the identity of the speaker [2]. This paper is mainly concern with the design and development of a speech corpus for multilingual speaker verification system.

Arunachal Pradesh is one of the linguistically richest and most diverse regions of Asia, being home to at least thirty and possibly as many as fifty distinct languages in addition to innumerable dialects and subdialects thereof[3]. The majority of languages indigenous to Arunachal Pradesh belong to the Tibeto-Burman language family. The majority of these in turn belong to a single branch of Tibeto-Burman, namely Tani. Almost all Tani languages are indigenous to central Arunachal Pradesh, including Nyishi, Adi, Galo, Apatani, Bangni, Tagin, Hills Miri, Bokar, Milang and many more. A handful of northern Tani languages are also spoken in small numbers in Tibet. Tani languages are noticeably characterized by an overall relative uniformity, suggesting relatively recent origin and dispersal within their present-day area of concentration. Most Tani languages are mutually intelligible with at least one other Tani language, meaning that the area constitutes a dialect chain, a characteristic which is commonly visible among European languages. However, only Apatani and Milang stand out as relatively unusual in the Tani context.

Although the state-of-the-art speech technology is developing in a full swing, yet speech resources remain scarce for minority languages of Arunachal Pradesh that have experienced a decline in the number of native speakers in the last few decades due to the growing popularity of Hindi among the educated section of people.

Performance of an Automatic Speaker Verification system is highly dependent on the speech database. There are lots of factors which affecting the performance of Automatic Speaker Verification system, some of them are recording conditions, modes, environment, devices, durations, speaker gender, age groups etc. Without knowing the recording condition, it is meaningless for expecting good result of an Automatic Speaker Verification system. The use of standard speech corpora for evaluation of ASR is the most crucial task in speech and speaker recognition system.

Manuscript received on January, 2013.

Utpal Bhattacharjee, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India.

Kshirod Sarmah, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India.

In the present study four native languages of Arunachal Pradesh, namely, Nyishi, Adi, Galo and Apatani have been considered for the development of the speech corpus along with English and Hindi. The speech corpus contain speech data from 200 speakers, 50 from each linguistic group, each group contain equal number of male and feamale speakers. Further, recoding has been done in English and Hind from each speaker

Some of important reasons why a speech corpus is needed have been given below:

a) A corpus with spoken audio data can expose language content, which is appropriate to the learner's geographical surroundings.
b) The practicalities of the characteristics of connected speech, which cannot be shown in written corpus texts, can be highlighted through the use of a spoken (speech) corpus, which benefits the learner's receptive skills.
c) Unlike text corpora, a speech corpus can reveal prosodic information and also shows preferred pronunciation styles of the chosen socio-cultural model.
d) This tool–which will satisfy the needs of teachers, learners and researchers – will link digitally recorded, natural, native-to-native speech so that each transcript segment will give access to its associated sound file.

II. A SURVEY ON STANDARD SPEECH CORPUS

The use of standard corpora for training and testing of speech and speaker recognition system is one of the major factors behind the rapid progress in automatic speech processing research, particularly in speech and speaker recognition in the last 10 years. There are lots of publicly available standard corpora for speaker recognition. The standard organizations like the Linguistic Data Consortium (LDC) [5], the European Language Resources Association (ELRA) [6] and the Oregon Graduate Institute (OGI) [7] are the main supplier of the standard corpora. However, most of the speech corpora were developed in American and Western context. These corpora are not suitable for speech recognition research in Asian context. A brief description of the major corpora have been given below[1,4,8]:

a) **TIMIT (LDC):** TIMIT (Texas Instruments Massachusetts Institute of Technology) database allows identification to be done under almost ideal conditions. The TIMIT database consists of 630 speakers, 70% male and 30% female from 10 different dialect regions in America. Each speaker has approximately 30 seconds of speech spread over ten utterances. The speech was recorded using a high quality microphone in a sound proof booth at a sampling frequency of 16 kHz, with no session interval between recordings. The TIMIT corpus was designed to provide speech data for aco ustic-phonetic studies and for developing and evaluating automatic speech recognition systems. Since it was one of the first corpora available with a large number of speakers.
b) **SIVA (ELRA):** The Italian speech corpus Speaker Identification and Verification Archives (SIVA) is composed of more than 2,000 calls collected over the public switched telephone network. The SIVA corpus consists of male and female users and male and female impostors. The database consist of 671 speakers where 335 male and 336 female. The number of sessions per speaker varies from 1 to 26. The type of speech includes prompted words, digits, short questions and read text. Various types of telephone handset microphones were used for recoding. The recoding has been done using PSTN channels. Home and office environment has been considered for recording.
c) **PolyVar (ELRA):** PolyVar is a speaker verification corpus comprised of native and non-native speakers of French, mainly from Switzerland. It consists of read and spontaneous speech in Swiss and French amounting to 160 hours of speech. The database consist of 143 speakers where 85 male and 58 female. Number of recording sessions each speaker participates varies from 1 to 229 resulting 3600 total recording sessions. The type of speech considered for recoding was read and prompted digits, word, sentences, questions and spontaneous speech. Microphones used for recoding are various telephone handsets microphone. Recoding has been done using PSTN channel (possibly ISDN). Acoustic environment considered for recoding was typical home/office.
d) **POLYCOST (ELRA):** The POLYCOST corpus was collected under the COST 250 European project for speaker verification. Most of the speech is non-native English with some speech in speaker's native tongue covering 13 European countries. The number of speakers considered for recoding was 133, where 74 male and 59 female. Each speaker is recorded for at least more than 5 sessions. Fixed and prompted digit strings, read sentences, free monologue have been considered for recoding. The recoding has been done using different handset microphones and digital ISDN channel. The acoustic environment recoding was typical home/office evaluation.
e) **YOHO (LDC):** The YOHO corpus is designed to support text-dependent speaker verification evaluation for Government secure access applications. A high-quality telephone handset (Shure XTH-383) was used to collect the speech; however, the speech was not passed through a telephone channel. YOHO was recorded in a fairly quiet office environment with low-level office noise, fan noise, and occasional pages over a public address system. The YOHO Database consists of 138 speakers, 108 of them male and 30 female. The data was collected over a three month period, with approximately 3 day verification intervals. The speech data consists of a series of combination-lock phrases, for example 24-52-78. There are 4 enrollment sessions per speaker, each containing 24 phrases and also contains 10 verification sessions per speakers, each containing 4 phrases. The data was recorded at 8 KHz with a 3.8 KHz bandwidth at 16 bits per sample.
f) **Switchboard I-II Including NIST Evaluation Subsets (LDC):** The Switchboard corpus represents one of the largest collections of conversational, telephone speech recordings available. There are two main Switchboard corpora (I and II), two phases of Switchboard-II and several subsets of Switchboard I-II used to create the NIST speaker recognition evaluation corpora. The Switchboard-I and Switchboard-II consist of 543 and 657 speakers respectively. Approximately 50% of the speakers were male and 50% female. The number of sessions varies from 1 to 25. Each session consist of 5 minutes conversation speech.

g) Different telephone handset microphones were used for creating the database. PSTN channel has been used for the recoding purpose. The acoustic environment considered was typical home and office environment.

h) **Speaker Recognition Corpus (OGI):** The Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology had collected a large speech database for speaker recognition research. The initial release of the corpus contains approximately 100 speakers consisting of 47 male and 53 female speakers calling from different telephone environments and at different times. Each speaker participates in 12 sessions. The speech data consist of prompted phrases, digits and prompted monologue. Different telephone handset microphone and PSTN channel has been used for recoding the speech signals. The acoustic environment considered was typical home/office environment.

i) **ANDOSL**: The ANDOSL (Australian National Database of Spoken Language) is a speech database jointly developed by the Australian National University, the University of Sydney, Macquarie University and the National Acoustic Laboratories, consisting of a few significant diverse phonological groups within Australia. The goal of ANDOSL was to represent as many significant speaker groups within the Australian population as possible. ANDOSL consists of 129 speakers 67 female and 62 male, from the three varieties of Australian English; Broad, General and Cultivated.

j) **Digit-SPL**: Digit-SPL is a multi-session database consisting of both male and female speakers. The database was developed at Griffith University in the early part of 2001 and consists of relatively clean speech (an average SNR of 41.6dB) from 68 males and 19 females, spoken on three separate sessions. The sessions are separated by approximately 4-8 weeks. All three session contain ten utterances of two continuously spoken random sequences of five digit numbers, where each digit appeared only once per utterance. The first session contains an additional five repetitions of the isolated word set: “zero”, “one”, “two”, “three”, “four”, “five”, “six”, “seven”, “eight”, “nine”.

III. IDEAL CHARACTERISTICS OF A CORPUS

Using standard speech corpus for development and evaluation of speech and speaker recognition research has proven to be very valuable in promoting progress in this area. For the development of an ideal corpus for speaker verification research in multilingual environment, there should have lots of important criteria which are listed below[2,9,10]:

a) Corpus should contain equal numbers of males and females to perform the experiments exclusively for male or female or both.

b) A good corpus will therefore strive to include as many styles and registers of language as possible, including samples of spoken language.

c) Speaker of wide ranging age should be considered to study the effect of age on pronunciations.

d) Corpus should be designed for closed set and open set ASR experiments.

e) Depth and breadth of the coverage- we must have enough speech from enough speakers in order to validate an experiment under study.

f) Data for a multilingual corpora should be equally collected from all possible phonologically distinct dialect zones in order to track all possible variation in pronunciation in that particular language.

g) Speaker of varying educational background should be considered in order to keep track of the effects of speech effort level and speaking rate, intelligence level, speaker’s experience, Lombard effect, etc.

h) Data should be collected with different types of microphones, transmission channels and medium such as mobile phones, laptop, fixed mounted headphone etc. in order to track the effect of devices variability on ASV system more realistic.

i) Data should be collected with multiple training and testing session in order to keep tract the effect of intersession variability on ASV performance.

j) An ideal corpus should be designed to evaluate a system’s performance against mimicry such as those based on physiological characteristics.

k) In a multilingual country like India, data should be collected from each of speaker in various languages for multilingual experiment.

l) Text material that given to speaker should be phonetically sound and balanced such that it contains all possible vowels, nasal consonants, fricatives and nasal-to-vowel co-articulations etc.

m) Data should be collected in wide range of acoustic environments such as home, office or telephone booth or road or cars or noisy conditions or workstations etc. in order to make ASV system more realistic.

n) On the issue of spontaneous speech, it is decided that the majority of the corpus should be read speech, for economic reason and the large amount of read speech will provide greater training benefits than smaller amount of spontaneous speech.

IV. DESCRIPTION OF THE SPEAKER VERIFICATION SPEECH CORPUS

A. Initial Step for the speech Corpora

The first step we followed in developing the ideal speech corpora is the selection of the textual sentences to be recorded from the speakers. The selected sentences should be minimal in number to save recording effort and at the same time have enough occurrences of each type of sounds to capture all types of co-articulation effects in the chosen language. So, in this work we selected 3-5 minutes duration textual sentences from some story books.

Some of the broad factors that can affect the performance of speaker verification system are[2]:

- Speech quality: types of microphones used, ambient noise levels, types of noise, compression of speech, etc.
- Speech modality: text-dependent or text–independent.
- Speech duration: amount to train and test data, temporal separation of training and testing data.
- Speaker population*:* number and similarity of speakers.

B. Experimental Setup

A typical experiment setup has been developed for recording the Speaker Recognition database.

The setup consists of a laptop with it good quality microphone, Fixed mounted microphone, a close talking microphone and a portable digital voice recorder. The two microphones are connected to two pc having Intel(R) Core ™ Duo CPU with speech processing software. Native Languages of Arunachal Pradesh considered for recording are given in table. 1.

Table 1. Native languages in the Database

Sl. No.	Native languages
1	Adi
2	Apatani
3	Galo
4	Nishi

The detail specification of the database is given below:

Table 2. Recording specification of the database

Number of Speakers	200 (Hindi and English), 50 Adi, 50 Apatani, 50 Galo and 50 Nyishi, (~50%M & ~50%F)
Number of sessions	4 for each language.
Intersession interval	2 weeks
Data types	Speech
Type of Speech	Read sentences
Sampling rate	16 KHz
Sampling format	Mono-channel, 16 bits resolution
Application	Text-independent (multilingual) ASV system
Speech Duration	3~5 minutes per speaker
Languages	Hindi, English and any one of the languages Adi, Apatani, Galo and Nishi which must be the mother tongue of the informant
Training segments	90s,120s,150s
Testing segments	10s, 20s,30s and 45s
Microphones	Fixed mounted, Close Talk, Laptop microphone and Portable Digital Recorder microphone
Acoustic environment	Laboratory
File Format	WAV PCM

C. Features of the Corpus

In this work we try to develop a new speech database (corpus) basically for the North-East Indian state Arunachal Pradesh Languages in order to research and development of some un-resourceful languages in this region. Some important characteristics of the corpus developed in this works are given below:

a) The two languages, viz., Hindi and English were used for common to all speakers of four native languages groups Adi, Apatani, Galo and Nyishi. So the data is completely phonetically independent.

b) Four languages, viz., Adi, Apatani,Galo and Nyishi were used for multilingual ASV system.

c) Data is recorded only in research laboratory environment, where the speaker may feel free to read the stories.

d) Speaker of ranging of ages (20-50 years) and variety of educational backgrounds (From the students of graduate and post graduate ,official staff as well as teaching staff of the University Campus) have been considered in the corpus.

e) Data are recorded in four different sessions having intervals 15-30 days.

f) The speakers were chosen such that all the diversities attributing to the gender, age and dialect are sufficiently captured.

g) The recordings were done with minimal background noise and mistakes in pronunciation, the errors that scripting in recording has been corrected manually by discarding the unsuitable parts of the utterance through manual listening.

V. CONCLUSION

In this paper, we state a simple methodology and a typical setup for developing a multilingual and multi-device speech corpora for Automatic Speaker Verification research for under resource languages of North-East Indian particularly the language of Arunachal Pradesh state of India, viz., Adi, Apatani, Galo and Nyishi. Various important characteristics of a speech corpus are discussed along with advantages and limitations of the other publicly available corpora and the features of the present corpora. The present corpora is also unique because of the fact that Arunachal Pradesh is the only region in whole of Asia where people from two major linguistic family namely Indo-European like Hindi, Bengali, Assamese, Nepali and Tibeto-Burman like Adi, Nyishi, Galo and Apatani are living together in the same geographic area and speak each others language fluently.

ACKNOWLEDGEMENT

This work has been supported by the ongoing project grant No. 12(12)/2009-ESD sponsored by the Department of Information Technology, Government of India.

REFERENCES

1. B.R. Wildermoth and K. K. Paliwal, GMM based speaker recognition on readily available databases, Proc Microelectronic Engineering Research Conf. 2003.
2. D.A. Reynolds, An overview of automatic speaker recognition technology, Acoustics, Speech, and Signal Processing (ICASSP), 2002, Vol. 4.
3. Wikipedia URL: http://en.wikipedia.org/wiki/ Arunachal_Pradesh
4. J.P.Campbell, Jr. and D.A. Reynolds, Corpora for the evaluation of speaker recognition systems, In Proceedings of International Conference on Acoustics, Speech and Signal Processing(ICASSP'99), 1999, Vol. 2, pp. 829–832.
5. Linguistic Data Consortium. URL: http:// www.ldc.upenn.edu/
6. European Lang Resources Assoc. http:// www.icp.grenet.fr/ELRA/
7. Oregon Graduate Institute URL: http:// cslu.cse.ogi.edu/
8. J.B. Millnar, J.P. VonWiller, J.M. Harrington and P.J. Dermody, The Austrakian national database of spoken language, in Proc. Inter. Conf. on Acoustics, Speech & Signal Processing (ICASSP'94), 1994, Vol. 1, pp. 97-101.
9. G. R. Doddington, CSR Corpus Development, In Proceedings of the workshop on Speech and Natural Language, 1992, pp. 363-366
10. H.A. Patil and T.k. Basu, Development of speech copora for speaker recognition research and evaluation in Indian languages, Int J Speech Technol, 2008, Vol. 11, pp. 17-32.

AUTHORS PROFIEL

Utpal Bhattacharjee received his Master of Computer Application (MCA) from Dibrugarh University, India and Ph.D. from Gauhati University, India in the year 1999 and 2008 respectively. Currently he is working as an Associate Professor in the department of Computer Science and Engineering of Rajiv Gandhi University, India. His research interest is in the field of Speech Processing and Robust Speech/Speaker Recognition.

Kshirod Sarmah received his Master of Science (M.Sc.) in Computer Science from Gauhati University, India in the year 2004. Currently he is pursuing his Ph.D. in Computer Science from Rajiv Gandhi University, India. His research interest is in the field of Speech Processing and Robust Speaker Recognition.