# A Survey based Study of Indian Language Speech Database for Speaker Recognition

Vijeta Verma
Assistant Professor,
Electrical & Electronics Department
O. P. Jindal Inst. of Technology,
Raigarh, India

Tomesh Verma
Assistant Professor,
Electronics & Telecommunication Department
Parthivi College of Engineering & Management
Bhilai-3 (C.G.) , India

Vinita Sahu
Assistant Professor,
Electronics & Telecommunication Department
Rungta College of Engineering & Technology
Raipur, India

***Abstract*— Verbal communication the most important and accepted form of communication between individuals. Human beings have elongate been provoked to create computer that can recognize and utter like human. When the researcher tries to develop firm recognition system they necessitate certain earlier stored data i.e. database for particular recognition system. There are a variety of speech databases existing for European Language but exceptionally less for Indian Language. In this paper we talk about the various Speech Database build up in various Indian Languages for speech recognition system and Text to Speech System.**

**General Terms:- Speech Recognition, Speech Database, Natural Language Processing, Human Computer Interaction.**

***Keywords:- Speech Recognition, Speech Corpus, Database, Speech Recognition.***

## I. INTRODUCTION

Speech is the most prominent and natural form of communication among humans. Lots of languages are spoken in the world. The computers System which can understand the spoken language can be very useful in domains like agriculture, health care and government services. Most of the Information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so that digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages.

Speech technologies can play a very important role in development of applications for common people in a multi-lingual society such as India which has about 1652 dialects/native languages. While Hindi the National language of India is written in Devanagari script, there other 17 languages that are been recognized by the Indian constitution. Other recognized languages by Indian Constitution are: Assamese, Tamil, Malayalam, Gujarati, Telugu, Oriya, Urdu, Bengali, Sanskrit, Kashmiri, Sindhi, Punjabi, Konkani, Marathi, Manipuri, Kannada and Nepali.

This paper describes development of speech corpora / database for few Indian languages. The various application specific Speech database are mentioned in the Section 2. Section 3 describes the General purpose speech database. Section 4 describes the few on the work being carried out in Labs. Section 5 describes the Speech corpora collected by the Linguistic Data Consortium for Indian Languages (LDC-IL). Section 6 gives the comparison of the studied speech database and the conclusion and discussion is in section 7

## II. APPLICATION OF SPECIFIC SPEECH DATABASE

A Project sanctioned by the Technology Development for Indian Languages (TDIL) for the development of Speech Recognition system for agriculture purpose using cell phones and landline in Marathi Language is being carried out at TIFR (Mumbai) and IIT Bombay jointly. The speech data for the project is been collected from the speaker at TIFR Mumbai and IIT Bombay using two dedicated phone line. For the development of database two volunteers are been appointed by the TIFR and IIT Bombay. They visit the various districts of Maharashtra and Collect the Speech Sample by calling the dedicated phone line at TIFR and IIT Bombay. The speech database will consist of data recorded from approximately 1500 speakers. As the data is recorded using phone lines it is narrow band speech along with background noise so the volunteers also have digital voice recorders to collect the wide band speech simultaneously when the speaker speaks on the phone line [2].

A Speech Database of Hindi language for Automatic Speech Recognition system for Travel domain has been developed at C- DAC Noida. The database consists of training data collected from 30 female speakers in a noise free environment consisting of approximately 26 hours of speech recordings. Total

8,567 sentences consisting 74,807 words were recorded by the speakers uniformly distributed over all age group from

17 to 60 years. The Recognition system was developed for the same recorded data and the recognition rate achieved for training data is 70.73% and for the test data is 60.66% [3].

A MIS (i.e. Mandi Information System) for retrieval of commodity price of market using mobile / telephone system is being developed at IIIT Hyderabad. The proposed MIS is in Telugu local language. The size of vocabulary of proposed recognition system is shown in the table 1.

TABLE 1. VOCABULARY SIZE USED IN MIS

| Word Category | Vocabulary Size |
|---|---|
| Commodity | 72 |
| Markets | 348 |
| Districts | 23 |

Speech data consisting of 17 hours of speech data was recorded from 96 speakers in noisy environment using mobile phones. A total of 500 words were recorded from each speaker. Approximately 15 hours of recorded speech data has been taken and used to build the acoustic model of ASR [4]. A speech to speech synthesis system for travel and Emergency services in Indian languages is developed at IIIT Hyderabad. The motivation for the said work was the problems faced by people who travel in India to see its rich cultural Heritage. The problem is when the people don‟t understand the native language were they visit, so a rapid development of Speech to Speech system in Telugu, Hindi and English has been done. Based on the collection of the possible usage scenarios, the broad domain of tourism and emergency services was divided into four different sub domains: 1) Local travel (D1) 2) Hotel and restaurant transactions (D2) 3) Tourism (D3) and 4) Emergency services (D4) for developing the speech synthesis system. The speech data was collected according to the said four domains. The Details of the sentences as per the domain and number of sentences is shown in table 2.
The speech databases developed for English, Telugu and Hindi was recorded from 15 different speakers

TABLE II. SPEECH CORPUS DETAILS

| | Number of Sentences | | |
|---|---|---|---|
| | English | Telugu | Hindi |
| D1 | 204 | 204 | -- |
| D2 | 206 | 206 | -- |
| D3 | 316 | 316 | -- |
| D4 | -- | 231 | 231 |

A Garhwali speech database is being developed for development of Automatic Speech Recognition system for Garhwali language at Government P.G. College, Rishikesh. A total number of 100 speakers consisting of 50 male and 50 female would be selected to speak the selected words or sentences. All speakers are from different district of Uttarakhand i.e. out of 13districts of Uttarakhand. They have considered Tehri Garhwal, Pauri Garhwal, Chamoli, Rudraprayag and Uttarakashi districts of Uttarakhand for recording the speech. In these districts of Uttarakhand Garhwali is spoken quite frequently. For developing the speech database a text corpus consisting 11,188 isolated Garhwali tokens/words has been prepared. For recoding the speech data PRAAT would be used. The speech recording would be done in the lab in noisy environment which would be helpful for the development of the robust speech recognition system [6].

## III. GENERAL PURPOSE SPEECH DATABASE

A Large Vocabulary Continuous Speech Database is developed at IIIT Hyderabad with coordination of HP Labs Bangalore. The developed database is in three different languages i.e. Marathi, Tamil and Telugu. The speech data was recorded using Mobile and Landline. In all 559 speakers participated for recording speech in all three different languages. The speakers who participated in recording procedure were from different age groups. The Speech data was collected from the native speakers of the language. Mobile phones and landlines were used to record the speech data from the speakers. The recorded speech consists of background noise and disturbance caused due to use of phone line [1].

A Punjabi language Speech Database has been developed for Text to Speech synthesis system at Punjabi University. The syllables are designed for Text to Speech Synthesis system because the researchers have selected syllables as the basic unit of concatenation. This Punjabi language speech database consists of 3,312 syllables. These syllables were adopted after analysis of total possible syllables of Punjabi corpus which was having nearly around four million words; it is found 9,317 were valid syllables with in which 3312 syllables were selected. A system for four Indian Languages is developed at Orissa University. For developing the speech corpora for the Text to Speech System in the said four languages native speakers were searched for all the four languages. The speakers were asked to read the text in the laboratory environment without any background noise. The text to speech synthesis system developed use the concatenation of syllables approach for the development of the Speech Database [8].

A General purpose, multi speaker, Continuous Speech Database has been developed for Hindi Language by the researchers of TIFR Mumbai and CDAC Noida. The Hindi Speech database is comprehensive enough to capture intra-speaker, phonetic, acoustic and inter speaker variability in Hindi Speech. This database consists of sets of 10 phonetically rich Hindi sentences spoken by 100 Native speakers of Hindi language. The speech data was digitally recorded using two microphones in a Noise free environment. Each speaker was asked to read the 10 sentences consisting 2 parts. The first part consists of two „Dialect‟ sentences which preferably covers the maximum phonemes of Hindi language. Every speaker was asked to speak these two sentences. The second part consisted of 8 sentences which covered maximum possible phonetic context. Though this continuous speech database was developed for training speech recognition system for Hindi language, it has been designed and developed in such a manner that is can also be used in tasks such as speaker recognition, study of acoustic-phonetic correlation of the language [9].

A General purpose speech database has been developed of Hindi, Telugu, Tamil, and Kannada from broadcasted news bulletin at IIT Kharagpur. The total database for the four languages is of 17.5 hours. Total durations of speech in Kannada, Hindi, Tamil, Telugu are 4.5 h, 3.5 h, 4.5 h, 5 h respectively. For Hindi Language data was recorded of 19 speakers (6 Male, 13 Females), for Telugu 20 Speakers (11 Male, 9 Females), for Tamil 33 Speakers (10 Male, 23 Females) and for Kannada 20 Speakers (12 Male, 8 Females). In each said languages these news bulletins were read by male and female speakers. As the speech database developed is of broadcast news the recording is done in the studio in a noise free environment [10].

A Text to Speech Synthesis for Konkani Language has been developed at Rajarambapu Institute of Technology Sakharale, Islampur, and Maharashtra. For the development of Text to Speech Synthesis a limited vocabulary speech database has been developed. The said database contains speech data recorded for more 1000 thousand Konkani commonly used words. Students were asked to take part as speaker for recording the speech data in their voice using standard microphone and a computer in the laboratory. The developed speech database consists of around 3,000 wave files consisting of Characters, Vowels, half Characters and Barakhadi [11]. A Speech database has been developed for developing a Text to Speech Synthesis system in Kannada Language at Mysore. The basic entity selected for the speech synthesis in this project was phonemes. This speech database consists of total 1,605 phonemes. The phonemes were recorded using the utility tool PRAAT on Windows Operating System platform [12].

At KIIT, Bhubaneswar a project for Mobile Text and Speech database collection in Hindi and Indian Spoken English has been completed. The Project was sponsored by Nokia Research Centre, China. The speech data was collected using 13 prompt sheets containing 630 phonetically rich sentences in each language prepared after collecting text messages in Hindi and Indian Spoken English. The collected text corpus for Hindi and English consists of 42,801 and 33,963 of unique Words respectively. The Speech data was recorded from 100 speakers for both the language each. The Speech data was recorded using 3 channels (i.e. mobile phone, Omi directional microphone and cardioid microphone) simultaneously at a sampling frequency of 16,000 Hz. The developed speech database consists 60% female voice recording and 40% male voice recording [13].

## IV. Various Works/ Project Going On For Speech Application Development

At Anna University project for Large Vocabulary Speech Recognition system and development of Language models for Tamil and Telugu Speech Recognition system is going on. At IBM Research Lab India a Telephone based Speech Recognition system for Hindi is being carried out. At CDAC Pune development of Speech to Text System for Hindi (i.e. Shrut-Lekhan) a prototype s ystem is being developed. At HP Labs India Speech Recognition for various Indian languages is going on. At CDAC Kolkata development of lexically driven Bengali Speech Recognition system is being carried out. Using the Wire or Wireless communication development of Speech based access for agricultural commodity is being carried out in 6 different Indian languages in the first phase. The work for Hindi is carried out at IIT Kanpur, for Assamese at IIT Guwahati, for Bengali at CDAC Kolkata, for Marathi at TIFR and IIT Mumbai (Combined), for Telugu at IIIT Hyderabad and for Tamil at IIT Chennai [14].

## V. Speech Corpora Collected By The Ldc-Il

The Linguistic Data Consortium for Indian Languages (LDC-IL) is the Consortium established after a long persuasion for developing a similar activity like Linguistic Data Consortium (LDC) at the University of Pennsylvania. The services of LDC-IL are been hosted and Managed by CIIL Mysore. It is also supported by the Central Government India. The LDC-IL will be responsible to create the database but will also provide forum for the researchers all over the world to develop speech application using the collected data in various domains.

The LDC-IL has collected Speech databases in various Indian Languages. The table 3 shows the Speech corpus collected by LDC- IL in hours [16].

## VI. Comparisons

The overall paper describes the speech database that are been developed for speech recognition system, text to speech synthesis system in some Indian languages. The developed speech databases are either for general purpose application or for task specific application.

In section 2 and 3 we have described briefly the various collected speech databases. The collected speech databases are compared with that of the instruments used for recordings, number of speakers, language, type of speech, the recording environment, language in which database is created and the application of the database. The table 4 shows the basis on which we have compared these different speech databases.

When we compare all the 13 databases that are studied we Observed that only 5 databases are been collected in noisy environment and 8 databases are recorded in Noise free or controlled environment. It shows that some of the databases are recorded using mobile phones or landline in such databases the speech data recorded is narrow band speech and many time the information may not be recorded because of the disturbance in the network or the phone line. It was observed that the speech databases that are been developed are for the Text to Speech Synthesis for which the database consists of phonemes or syllables. The Linguistic Data Consortium for Indian Languages (LDC-IL) has collected a huge speech corpus in different Indian languages and they are ready to distribute the database to the researchers for developing the application.

The databases that are been developed for Text to Speech synthesis system generally consists phonemes or syllables as the basic concatenative unit. Such types of databases are not that effective for continuous speech recognition system. Lots of work is been carried out for Hindi and Telugu languages. Little work is been done for other Indian languages. Researchers should try to work for other Indian languages so that language technologies can be developed in all the Indian Languages.

## VII. Conclusions

In this paper we have discussed some of the speech databases developed in different Indian languages for various applications. We have also mentioned the various research projects that are going for development of speech recognition system and text to speech synthesis system. Through this review we have found that the majority of the work is been done for the Indian languages like Hindi, Tamil, Telugu and Bengali. Little work has been done or is going on for the Marathi Language. The systems that are been developed are in preliminary stage. The accuracy of these developed systems is less and they are just developing the recognition system on trail basis at the initial phase but not a complete recognition system is been developed yet.

The research that has been carried out is mostly for text to speech synthesis which uses phoneme/syllables concatenation or isolated words. The need for today's speech application is to work on the Continuous speech. The research that has been today's speech application is to work on the Continuous speech.

The researchers should try to develop the speech database in the noisy environment which will help to develop noise robust speech recognition systems which will be useful in the real life scenarios and will work efficiently. This study will help the researchers to know the work that has been completed and the work that is been carried out at the different research institute and universities. After studying the developed Indian speech databases we have been motivated to develop a continuous speech database in Marathi language for agriculture database. We will try to cover the maximum phonetic variation and different environment while recording the speech.

TABLE III. Speech Corpus Details

| | | |
|---|---|---|
| 1. | Assamese | 105:52:37 |
| 2. | Bengali | 138:18:47 |
| 3. | Bodo | 114:38:55 |
| 4. | Dogri | 58:12:49 |
| 5. | Gujarati | 146:23:04 |
| 6. | Hindi | 163;25:47 |
| 7. | Indian English Bengali | 34:12:57 |
| 8. | Indian English Gujarati (MP3) | 21:40:00 |
| 9. | Indian English Kannada | 37:01:33 |
| 10. | Kannada | 137:53:28 |
| 11. | Kashmiri | 44:59:07 |
| 12. | Konkani | 205:01:48 |
| 13. | Maithili | 43:33:42 |
| 14. | Malayalam | 105:47:05 |
| 15. | Manipuri | 107:10:30 |
| 16. | Marathi | 168:13:50 |
| 17. | Nepali | 145:04:46 |
| 18. | Oriya | 45:10:25 |
| 19. | Punjabi | 71:55:56 |
| 20. | Tamil | 87:03:24 |
| 21. | Telugu | 50:51:36 |
| 22. | Urdu | 81:06:25 |

## References

[1]. Gopalakrishna Anumanchipalli, Chitturi R., Joshi S., Kumar R., Singh S., R. Sitaram, S. Kishore. 2005. " Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems", International Conference on Speech and Computer (SPECOM), Patras, Greece. Languages. In Proceeding of Joint Workshop on Hands- free Speech Communication and Microphone Arrays (HSCMA), Edinburg, Scotland

[2]. Tejas Godambe and Samudravijaya K., "Speech Data Acquisition for voice based Agricultural Information Retrieval" Proceeding of 39th All India DLA Conference, Punjabi University, Patiala, India

[3]. Arora S., Saxena B., Arora K., Agarwal S., "Hindi ASR for Travel Domain", Proceedings of OCOCOSDA'10, Kathmandu, Nepal.

[4]. Mantena G., Rajendran S. , Rambabu , Suryakanth V., Gangashetty, B., Yegnanarayana K.. "A Speech-Based Conversation System for Accessing Agriculture Commodity Prices in Indian"

[5]. Anandaswarup , Karthika, Nagaswetha, Narne, Vinay B., Mrudula Poornima, Patil, Raju, Snehata, "Rapid Development of Speech to Speech Systems for Tourism and Emergency Services in Indian Languages". International Conference on Services in Emerging Markets, Hyderabad, India. Upadhyay, Riyal "Garhwali Speech Database", O-COCOSDA'10, Kathmandu, Nepal.

[6]. Singh P., Lehal G.. "Text-To- Speech Synthesis System for Punjabi Language", International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain.

[7]. S. Mohanty, "Syllable Based Indian Language Text To Speech System", International Journal of Advances in Engineering & Technology, Vol.1, Is. 2.

[8]. Samudravijaya, Rao and Agarwal., "Hindi Speech Database". Sixth International Conference on Spoken Language Processing, Beijing,

[9]. K. Rao, "Application Prosody model for Developing speech system", International Journal of Speech Technology, Vol. 11, Elseveir.

[10]. Borkar, Patil, "Text To Speech System For Konkani Language". W3C Workshop on Internationalizing the Speech Synthesis Markup Language III

[11]. Ravi, Patilkulkarni, “A Novel Approach to Develop Speech Database for Kannada Text-to-Speech System”, Int. J. on Recent Trends in Engineering & Technology, Vol. 05, No. 01,

[12]. Agrawal, Sinha, Singh, Olsen. “Development of Text and Speech Database for Hindi and Indian English specific to Mobile Communication Environment”. International Conference on The Language Resources and Evaluation Conference, LREC, Istanbul,

[13]. Agrawal S .“Recent Developments in Speech Corpora in Indian Languages: Country Report of India”. O-COCOSDA, Kathmandu,

[14]. Chalapathy, Rajput, Verma. “A Large Vocabulary Continuous Speech Recognition system for Hindi”. the National conference on Communications, Mumbai, pp. 366 -370.