Indian Language Speech Database: A Review

Pukhraj P. Shrishrimal
Dept. of CS and IT
Dr. B. A. M. University,
Aurangabad-431004, India

Ratnadeep R. Deshmukh
Dept. of CS and IT
Dr. B. A. M. University,
Aurangabad-431004, India

Vishal B. Waghmare
Dept. of CS and IT
Dr. B. A. M. University,
Aurangabad-431004, India

ABSTRACT
Speech is the most prominent and natural form of communication between humans. Human beings have long been motivated to create computer that can understand and talk like human. When the research tries to develop certain recognition system they require certain previously stored data i.e. database for respective recognition system. There are various speech databases available for European Language but very less for Indian Language. In this paper we discuss the various Speech Database developed in different Indian Languages for speech recognition system & Text to Speech System.

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

Keywords
Speech Recognition, Speech Corpus, Database, Speech Recognition.

1. INTRODUCTION
Speech is the most prominent and natural form of communication between humans. There are various spoken languages throughout the world. Communication among the human being is dominated by spoken language, therefore it is natural for people to expect speech interfaces with computer.

Speech has potential of being used as a mode of interaction with computer. Human beings have long been motivated to create computer that can understand and talk like human. In this direction, researchers have tried to develop system for analysis and classification of the speech signals. Since, 1960s computer scientists have been researching ways and means to make computer record, interpret and understand human speech.

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 constitution of India. The other languages recognized by Indian Constitution are: 1) Assamese 2) Tamil 3) Malayalam 4) Gujarati 5) Telugu 6) Oriya 7) Urdu 8) Bengali 9) Sanskrit 10) Kashmiri 11) Sindhi 12) Punjabi 13) Konkani 14) Marathi 15) Manipuri 16) Kannada and 17) Nepali [1].

The amount of work for Indian languages in Speech domain has not yet reached to a critical level to be used as real communication tool, as that in other languages of developed countries. Few attempts to develop speech recognition system had been attempted by HP Labs India and IBM research lab [1, 15]. However, there is lot of scope to develop language technology systems using Indian languages which are of different variations. To achieve such ambitious goals, the collection of standard speech databases is prerequisite.

This paper describes the 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 at the various Intuitions, Universities and Research 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.

2. APPLICATION 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 language. The vocabulary size of proposed system is shown in the table 1.

Table 1. Vocabulary size used in Mandi Information System

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.

Table 2. Speech Corpus Details

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

The speech databases developed for English, Telugu and Hindi was recorded from 15 different speakers. All the recordings were done using a laptop and a standard microphone in a room in noise free environment. [5].

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].

3. 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 Department of Computer Science, Punjabi University, Patiala. The syllables were considered for developing said speech database 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 which account for more than 99% of commutative percentage frequency in the selected corpus. These syllables were selected after analyzing total possible syllables of Punjabi corpus which was having nearly 2, 33,009 unique and more than four million words; out of which 9,317 were valid syllables from which 3312 syllables were selected. The selected syllables were recorded from a speaker using standard microphone in the studio environment. [7].

A Text to Speech synthesis System for four Indian Languages Hindi, Odiya, Bengali and Telugu has been developed at Department of Computer Science and Application, Utkal University, Bhubaneswar. 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 phonetic, acoustic, intra-speaker and inter speaker variability's 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. This database is used for developing the prosody models for Speech recognition, Speech Synthesis, Speaker Recognition and Language Identification Application. The total database for the four languages is of 17.5 hours. Total durations of speech in Hindi, Telugu, Tamil and Kannada are 3.5 h, 4.5 h, 5 h and 4.5 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, 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 Vowels, Characters, Barakhadi and half Characters [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. The sampling frequency used for recording the speech was 16,000 Hz. The recording was done using the standard microphone in lab. The recorded phonemes include vowels, semi vowels, stops, fricatives, nasals etc. [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]

4. VARIOUS WORKS/PROJECT GOING ON FOR SPEECH APPLICATION DEVELOPMENT

There are various research projects / work going in India related to speech Recognition, speaker recognition, Speech Synthesis and Machine Translation.

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 system is being developed. At HP Labs India Speech Recognition for various Indian languages is going on. They are working on development of Large Vocabulary Continuous Speech Recognition system. [14]

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 Madras. [14]

5. 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].

Table 3. Speech Corpus Details

Sr. No.	Language	Hours
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

6. COMPARISON

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 concatinative unit. Such types of databases are not that effective for continuous speech recognition system. The maximum 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.

Table 4. Comparison of the Databases

Sr. No.	Database Developed by	Recording Environment	No. of Speakers	Recording Device Used	Application of Database	Language
1.	TIFR Mumbai and IIT Bombay	Noisy Environment	1500	Cell Phone & Voice Recorders	Speech Recognition System for Agriculture Purpose	Marathi
2.	C-DAC Noida	Noise Free Environment	30	Standard Mics	Speech Recognition System for Travel Domain	Hindi
3.	IIIT Hyderabad	Noisy Environment	96	Mobile Phones	Speech Recognition System for Agricultural commodity Price Enquiry	Telugu
4.	IIIT Hyderabad	Noise Free Environment	15	Standard Microphone and Laptop	Travel and Emergency Services	Telugu, Hindi & English
5.	Government P.G. College, Rishikesh	Noisy Environment	100	Standard Microphones	Speech Recognition System	Garhwali
6.	Punjabi University, Patiala	Studio Environment	1	Standard Microphone	Text to Speech Synthesis System	Punjabi
7.	Utkal University, Bhubaneswar	laboratory environment	Not Known	Noise Cancellation Microphone	Text to Speech Synthesis System	Hindi, Odiya, Bengali & Telugu
8.	TIFR Mumbai and C-DAC Noida	Noise Free Environment	100	Standard Microphone	General Purpose	Hindi
9.	IIT Kharagpur	Studio Environment	92	Standard Microphone	General Purpose	Hindi, Telugu, Tamil, & Kannada
10.	Islampur, Maharashtra	Laboratory Environment	Not Known	Standard Microphone	Text to Speech Synthesis System	Konkani (Goan)
11.	SJ College of Engineering, Mysore	Laboratory Environment	Not Known	Standard Microphone	Text to Speech Synthesis System	Kannada
12.	KIIT, Bhubaneswar	office environment	200	Cellphone, Omni directional & cardioid Microphone	Mobile based speech recognition	Hindi & Indian Spoken English
13.	IIIT Hyderabad and HP Labs Bangalore	Noisy Environment	559	Mobile Phone and Landline	General Purpose	Marathi, Tamil & Telugu

7. CONCLUSION

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 carried out is mostly for text to speech synthesis uses phoneme/syllables concatenation or isolated words. The need for 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.

8. ACKNOWLEDGEMENT

The authors would like to thank the University Authorities for providing the infrastructure to carry out the research. This work is supported by University Grants Commission.

9. REFERENCES

[1] Gopalakrishna Anumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh, R. N. V. Sitaram, S. P. Kishore. 2005. Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems. In Proceedings of International Conference on Speech and Computer (SPECOM), Patras, Greece.

[2] Tejas Godambe and Samudravijaya K. 2011. Speech Data Acquisition for voice based Agricultural Information Retrieval. In Proceeding of 39th All India DLA Conference, Punjabi University, Patiala, India

[3] Sunita Arora, Babita Saxena, Karunesh Arora, S S Agarwal. 2010. Hindi ASR for Travel Domain. In Proceedings of OCOCOSDA 2010, Kathmandu, Nepal.

[4] Gautam Varma Mantena, S. Rajendran, B. Rambabu, Suryakanth V. Gangashetty, B. Yegnanarayana, Kishore Prahallad. 2011. A Speech-Based Conversation System for Accessing Agriculture Commodity Prices in Indian Languages. In Proceeding of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), Edinburg, Scotland.

[5] Anandaswarup V, Karthika M, Nagaswetha G, PK Narne, VV Vinay Babu, Mrudula K, Poornima T, RR Patil, CMS Raju, Snehata T, Azharuddin S, Abhilash B, P Raju, GSC Prasad, Sriram A, E Veera Raghavendra, Sachin Joshi, Vamshi Ambatiy and Kishore S Prahallad. 2010. Rapid Development of Speech to Speech Systems for Tourism and Emergency Services in Indian Languages. In Proceeding of International Conference on Services in Emerging Markets, Hyderabad, India.

[6] R. K. Upadhyay and M. K. Riyal. 2010. Garhwali Speech Database. In Proceedings of O-COCOSDA 2010, Kathmandu, Nepal.

[7] Parminder Singh, Gurpreet Singh Lehal. 2006. Text-To-Speech Synthesis System for Punjabi Language. In Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain.

[8] Sanghamitra Mohanty, "Syllable Based Indian Language Text To Speech System", International Journal of Advances in Engineering & Technology, 2011. Vol. 1,Issue 2.

[9] Samudravijaya K., P. V. S. Rao and S. S. Agarwal. 2000. Hindi Speech Database. In Proceedings of Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China.

[10] K. Sreenivas Rao, "Application Prosody model for Developing speech system", International Journal of Speech Technology, 2011, Vol. 11 in Elseveir

[11] Sangam P. Borkar and Prof. S. P. Patil. 2007. Text To Speech System For Konkani (Goan) Language. In Proceedings of W3C Workshop on Internationalizing the Speech Synthesis Markup Language III — Agenda

[12] D. J. Ravi and Sudarshan Patilkulkarni, "A Novel Approach to Develop Speech Database for Kannada Text-to-Speech System", Int. J. on Recent Trends in Engineering & Technology, 2011, Vol. 05, No. 01, in ACEEE.

[13] Shyam Agrawal, Shweta Sinha, Pooja Singh, Jesper Olsen. 2012. Development of Text and Speech Database for Hindi and Indian Enlish specific to Mobile Communication Environment. In Proceeding of International Conference on The Language Resources and Evaluation Conference, LREC, Istanbul, Turkey.

[14] Agrawal S. S. 2010 Recent Developments in Speech Corpora in Indian Languages: Country Report of India. O-COCOSDA, Kathmandu, Nepal.

[15] Chalapathy Neti, Nitendra Rajput, Ashish Verma. 2002 A Large Vocabulary Continuous Speech Recognition system for Hindi. In Proceedings of the National conference on Communications, Mumbai, pp. 366-370.

[16] Size of Speech Corpora (As on Dec 2011) , Available at: http://www.ldcil.org/resourcesSpeechCorp.aspx