Int. J. Advanced Networking and Applications
Volume: 06 Issue: 06 Pages: 2556-2558 (2015) ISSN: 0975-0290

A Review on Speech Corpus Development for Automatic Speech Recognition in Indian Languages

Cini kurian
Department of computer Science, Al-Ameen college, Edathala,Aluva, Kerala
cinikurian@gmail.com

ABSTRACT

Corpus development gained much attention due to recent statistics based natural language processing. It has new applications in Language Technology, linguistic research, language education and information exchange. Corpus based Language research has an innovative outlook which will discard the aged linguistic theories. Speech corpus is the essential resources for building a speech recognizer. One of the main challenges faced by speech scientist is the unavailability of these resources. Very fewer efforts have been made in Indian languages to make these resources available to public compared to English. In this paper we review the efforts made in Indian languages for developing speech corpus for automatic speech recognition.

Keywords - **Speech corpus, Indian languages**

Date of Submission: August 28, 2014 Date of Acceptance: December 29, 2014

I. INTRODUCTION

In modern linguistic, Corpus is the machine readable form of the large collection of structured text in written or spoken form[1]. If corpora can give some linguistic information it is called annotated Corpora. In English and in other languages many path breaking researches have been done and many pioneering computer based systems have been developed using language corpora. Importance of language corpora is recognized in many countries. However, as far as India is concerned, using corpora in language and NLP research is a time taking process as it is difficult to capture fancy of Indian linguistics because of its diversity. Importance of language corpora is recognized in many countries. 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 a prerequisite. This paper review the development of speech corpus in Indian languages

II.SURVEY ON SPEECH CORPUS OF INDIAN LANGUAGES

Speech corpus has been collected in Marathi language at TIFR (Mumbai) and IIT Bombay jointly as a part of a project sanctioned by the Technology Development for Indian Languages (TDIL) for the development of speech recognition system for agriculture purpose. The speech data for the project was collected from the speaker at TIFR Mumbai and IIT Bombay using two dedicated phone line. For the development of database two volunteers were appointed by the TIFR and IIT Bombay. They visited the various districts of Maharashtra and Collect the speech sample by calling the dedicated phone line at TIFR and IIT Bombay. The speech database consists of data recorded from approximately 1500 speakers. As the data was 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 was 70.73% and that for the test data was 60.66% [3].

A MIS (i.e. Mandi Information System) for retrieval of commodity price of market using mobile/telephone system has been developed at IIIT Hyderabad. The proposed MIS was in Telugu language. Speech corpus consisting of 17 hours of speech data 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 collected 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 speech databases developed include English, Telugu and Hindi speech corpus 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 developed for development of automatic speech recognition system for Garhwali language at Government P.G. College, Rishikesh. Total number of 100 speakers consisting of 50 male and 50 female have been selected to speak the selected words or sentences. All speakers were from

Volume: 06 Issue: 06 Pages: 2556-2558 (2015) ISSN: 0975-0290

different district of Uttarakhand. The speech recording was done in the lab in noisy environment which would be helpful for the development of the robust speech recognition system [6].

A large vocabulary continuous speech database is developed at IIIT Hyderabad with coordination of HP Labs Bangalore. The developed database was in three different languages i.e. Marathi, Tamil and Telugu. The speech data was recorded using mobile and landline. A total of 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 is caused due to use of phone line [7].

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 the 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 [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 variabilities 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 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].

Another general purpose speech database has been developed in Hindi, Telugu, Tamil, and Kannada from broadcasted news bulletin at IIT Kharagpur. This database was 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 from 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 language 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].

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 languages. The speech data was recorded using 3 channels (i.e. mobile phone, Omi directional microphone and cardioids microphone) simultaneously at a sampling frequency of 16,000 Hz. The developed speech database consists 60% female voice recording and 40% male voice recording [11].

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 than 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 and half Characters [12].

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 [13].

III. COLLECTIVE EFFORTS

India, as a multilingual country realized its prospects and DOE (Department of Electronics, Govt. of India) under the TDIL (Technology Development for Indian Languages)program has initiated some work on corpora development of all major Indian languages. Under this program, in association with CIIL(Central Institute of Indian Languages) , machine readable corpora for major

Volume: 06 Issue: 06 Pages: 2556-2558 (2015) ISSN: 0975-0290

languages has been developed . While comparing with British National Corpus (BNC) which contains data obtained from people on all walks of life, we are in the infancy stage. About 50 hours of annotated speech corpora for Hindi, Marati, Punjabi, Bengali, Assamese, and Manipuri languages have been developed by C-DAC (Centre for Development of Advanced Computing) Noida and C-DAC Calcutta. Corpus generation in India is facing several problems due to lack of a centralized authority (a consortium). Many organizations and institutes have collected corpus for their own research activities. However, these resources are not available to all groups of people working for corpus generation. TDIL and CIIL have taken some initiatives and put the data on the web and their contributions have been appreciated. But there should be national archive for Indian language corpora, so that all corpora will be systematically preserved, documented, distributed, accessed by the users.

The Linguistic Data Consortium (LDC), European Language Resources Association (ELRA), and The International Computer Archive of Modern and Medieval English (ICAME) are the good models for this.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 have been hosted and managed by CIIL Mysore. It is also supported by the Central Government of India. The LDC-IL is responsible to create the database and to 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 [168].

IV.CONCLUSION

Speech corpus applications have tremendous prospects in India since it is an essential component of speech recognition research. An attempt has been made through this paper to give a comprehensive survey on the development of corpus for automatic speech recognition in Indian languages.

REFERENCES

[1] Dash, N S and B B Chaudhuri. "Why do we need to develop corpora in Indian languages", *International Conference on SCALLA, Banglore, 2001*

[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 Handsfree Speech Communication and Microphone Arrays (HSCMA), Edinburg, Scotland.

[5] Anandaswarup V, Karthika M, Nagaswetha G, 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] 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.

[8] 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.

[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 Elsevier.

[11] 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.

[12] 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.

[13] 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.