Scientific Research Publishing

Open Journal of Applied Sciences, 2017, 7, 233-241
http://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Resources for Development of Hindi Speech Synthesis System: An Overview

Archana Balyan

Department of Electronics and Communication, Maharaja Surajmal Institute of Technology, Affiliated to GGSIPU, New Delhi, India
Email: archanabalyan@rediffmail.com

How to cite this paper: Balyan, A. (2017) Resources for Development of Hindi Speech Synthesis System: An Overview. *Open Journal of Applied Sciences*, **7**, 233-241.
https://doi.org/10.4236/ojapps.2017.76020

Received: May 1, 2017
Accepted: June 16, 2017
Published: June 19, 2017

Abstract

Most of the information in digital world is accessible to few who can read or understand a particular language. The speech corpus acquisition is an essential part of all spoken technology systems. The quality and the volume of speech data in corpus directly affect the accuracy of the system. However, there are a lot of scopes to develop speech technology system using Hindi language which is spoken primarily in India. To achieve such an ambitious goal, the collection of standard database is a prerequisite. This paper summarizes the Hindi corpus and lexical resources being developed by various organizations across the country.

Keywords

Speech, Database, Corpora, Lexicon, Speech Synthesis, Linguistics, Natural Language Processing

1. Introduction

The objective of speech data collection is to primarily build speech recognition and synthesis systems for Indian languages [1]. There is an ever-growing demand for customized and domain-specific voices for use in corpus based on synthesis systems. Hence, it is very important that good methods should be established for creating these databases. The high-quality audio data, and of large volume, is key to developing a high-quality speech synthesizer. In a country like India, where the literacy rate is low, Indian language speech interfaces can provide access to IT applications and services, through the Internet and/or telephones, to the masses. So that people in various semi-urban and rural parts of India will be able to use telephones and Internet to access a wide range of services and information on health, agriculture, travel, etc. However, for this to become a reality, computers should be able to accept speech input in the user's language and provide speech output. Also, in multilingual India, if speech technology is coupled with translation systems between the various Indian languages, services and information can be provided across languages more easily. Due to the lack of appropriate annotated speech databases in Indian languages, robust applications have not been developed. Efforts are being made by a selected set of Indian academic and research institutions in a consortium mode to build speech synthesis, speech recognition and machine translation systems in Indian languages. These efforts are primarily supported by the ministry of the information and communication technologies (MCIT), Govt. of India (GoI). The resources including speech and text corpora collected in these efforts abide by the copyright restrictions of the sponsor [2].

Hindi is an Indo-Aryan language with about 545 million speakers, 425 million of whom are native speakers. As per the eighth schedule of government of India, there are 22 official languages and it is one of the official languages of India and national language of the Federal Government of India. Hindi is spoken by a maximum number of people by about 41% of the population mostly in northern, central, western and eastern parts of the country [3]. This paper focuses on the Hindi resources being developed, which can be used for research in computational linguistics.

2. Hindi Text Encoding

The computer age in India began in 1955 with the installation of HEC-2M (Hollerith Electronic computer model-2M) a computer designed by A.D. Booth in England) at the Indian Statistical Institute (ISI) at Calcutta (now Kolkata) [4].

Code Standardization for Indic Scripts: A Survey

Various letters of the input text are recognized and converted into their respective codes. In 1978, India's DoE constituted a standardization committee, for designing codes for Indic scripts similar to ASCII. In 1982, first version of a 7-bit code, called ISSCCI-7 (Indian scripts Standard Code for information Interchange). In 1983, the first version of 8-bit code (ISCII-8) was released. A further modification was made in 1991, and the Bureau of Indian standards accepted ISCII-8 as national standard (IS 13194:1991). A newly formed Unicode consortium adopted 1998 version of ISCII-8 as the base for 16-bit Unicode for allocating codes to different Indian scripts [5]. With the advent of Unicode in 1990s, some online publications have switched to Unicode. A main on-line source of Hindi text in Unicode is Universal Word—Hindi dictionary [6] is being made at CFILT, IIT Bombay for the purpose of Machine Translation. The user can search the Hindi and English words and phrases. This lexicon also provides the grammatical, morphological and semantic attributes of the Hindi words. This version contains 36,111 Words and is good source of corpus. Encoding conversion may be required if data is acquired from other sources.

3. Corpora Development in Hindi Language

In modern linguistics, a corpus is the machine readable form of large collection of structured text in written or spoken form [7]. If corpora can give some linguistic information, it is called Annotated Corpora. It is as important a resource as any other in the field of language engineering. With the recent advancement in computer technology the availability of language corpora (by corpora we mean corpus) and its processing has become even easier and has opened many new areas of research in language processing. A corpus can be the best resource to study many different linguistic phenomena such as the spelling variations, morphological structure, and word sense analysis and how the language has evolved over the time and many more [8].

3.1. Development of Speech Corpora

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 8567 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% [9].

Another general purpose speech database in Hindi has been developed from Broadcasted news bulletin at IIT, Kharagpur. The total duration of speech in Hindi is 3.5 Hrs and was recorded for 19 speakers (6 Males and 13 Females). As the speech database is of broadcast, the recording is done in the studio in a Noise free environment [10].

The IIIT Hyderabad India developed speech databases at Speech and Vision Lab, for the purpose of building speech synthesis systems in Indian languages. This database consists of text and speech data in Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu. In the data base the text content collected from more than 10,000 Wikipedia articles. This database is collected from native speaker who are available in this campus. Each of these languages has several dialects. The speech data was recorded by a native speaker of the language. The recording was done in a studio environment using a standard headset microphone connected to a zoom handy recorder [11].

Two main sources for Mobile database at are developed at KIIT, Bhubneshwar and KIIT, Gurgaon. At KIIT, Gurgaon, a text corpus of 2 million words of natural messages in 12 different domains in Hindi and Indian English and a speech corpus of 100 speakers, each speaking 630 phonetically rich sentences, has been created. The speech utterances were recorded in 16 kHz through 3 recording channels: a mobile phone, a headset and a desktop mounted microphone. This project was sponsored by Nokia Research Centre China [11].

The Linguistic Data Consortium for Indian Languages (LDCIL) is the Consortium responsible for creating database and tools for collection of high quality database in various domains that can be used by researchers in developing speech technology systems. At LDCIL [12], a Hindi Speech Recognition database was collected in Uttar Pradesh and Bihar and contains the voices of 650 different native speaker who were selected according to age distribution (16 - 20, 21 - 50, 51+), Gender, Dialectical Regions and environment (home, office and public place). Each speaker record read a news text in a noisy environment through recorder having an inbuilt microphone. The recordings are in stereo recording and the extracted channel is also included in the specific files. It includes audio file, text file, NIST files which were saved as. ZIP Files. All the speech data are transcribed and labeled at the sentence level.

The purpose of developing the IIIT-H Indic speech databases is to have speech and text corpora made available in the public domain, without copyright restrictions for non-commercial and commercial use. A text 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. This application is needful for people who faced problem for travel in India to see its rich cultural Heritage. All the recordings were done using a laptop and a standard microphone in a room in noise free environment [13].

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 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 [14].

At KIIT, Bhubaneswar, a project for mobile text and speech database collection in Hindi 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 Hindi language after collecting text messages in Hindi. The collected text corpus for Hindi consists of 42,801 unique words respectively. The speech data was recorded from 100 speakers using 3 channels simultaneously at a sampling frequency 16 KHz. The developed speech database consists of 60% female voice recording and 40% male voice recording [14].

3.2. Development of Textual Corpora

EMILLE [15] Project (Enabling Minority Language Engineering), initiated by

Lanchester University, is one of the first initiatives taken to make Hindi corpus available for research and development of the language processing. The project has released 200,000 words of English text translated to Bengali, Guajarati, Hindi, Punjabi and Urdu creating a parallel corpus across these languages [16].

3.2.1. Indian Resources: Web Corpora for Indian Text

The Leipzig Corpora Collection (LCC) [17] has been collecting digital text material for more than 30 years. Over the last years, the established text acquisition and text processing tools are adopted to deal with Indian language to create and improve resources based on Indian text material. Corpora of this collection are typically grouped regarding the dimensions language, country of origin, text type (newspaper text, governmental text, generic Web material, religious texts etc.) and time of acquisition. Table 1 gives an introduction to currently available resources. It contains the number of sentences for Hindi languages and genres. The corpora are available via Web-based interfaces [7].

A main on-line source of Hindi text in Unicode is Universal Word—Hindi dictionary [18] is being made at Center for Indian Language Technology (CFILT), IIT Bombay for the purpose of Machine Translation. The user can search the Hindi and English words and phrases. This lexicon also provides the grammatical, morphological and semantic attributes of the Hindi words. This version contains 36,111 words and is good source of corpus. Encoding conversion may be required if data is acquired from other sources.

C-DAC Noida has created Gyan Nidhi Corpus, which is parallel in multiple Indian languages. GyanNidhi contains 1 million pages of digitized data in Unicode format which contains variety of data from books published by national book Trust, India, Sahitya Akadmi, Navjivan publications, Publication division, Shri Aurobindo Ashram as they publish books of various domains, in most of the Indian languages. Mahatma Gandhi Hindi International University has commenced a project "Hindi Samgraha" on databases and dialect mapping of Hindi [19].

3.2.2. Indian Resources: Web Corpora for Indian Text

A Lexical Resource, "Syntax and Morphology in Hindi and Urdu" [20] is a searchable database with entries for about 60 verbs, part of a larger database project which is in progress. Each entry has fields for information about verb attributes, which for a specific entry define important properties which are projected into a sentence. This lexical information constrains the possible sentences formed from this verb: for example, the number of arguments which verb takes their category and grammatical function, and the case forms which are required or possible. Lexica are as critical for development of language computing as Corpora. Two available manually compiled English-Hindi electronic dictionaries have been

Table 1. Amount of available resources in a number of sentences.

Language	News	Wikipedia	For comparison; EMILLE
Hindi	5,162,167	727,882	469,395

identified. First is the SHABDKOSH [21] and the second one is SHABDANJALI [22]. These two dictionaries have been merged automatically by replacing the duplicates. The merged English-Hindi dictionary contains approximately 90,872 unique entries. The positive and negative sentiment scores for the Hindi words are copied from their English SentiWordNet. The bilingual dictionary based translation process has resulted 22,708 Hindi entries [23]. In addition, English to Indian languages synsets are being developed under Project English to Indian Languages Machine Translation Systems (EILMT), a consortia project funded by Department of Information Technology (DIT), Government of India. For each language we have approximately 9966 synsets along with the English WordNet offset [23]. Hindi WordNet is a well structured and manually compiled resource and is being updated since last nine years. There is an available API [24] for accessing the Hindi WordNet [7].

3.2.3. Sentiment Lexicon for Hindi

Creation of linguistic data using SentiWordNet(s) for Indian languages are being developed using various approaches which can be used in areas of NLP too.

- WordNet(s) are available for Hindi (Jha *et al.*, 2001) [25].
- (Joshi *et al.*, 2010) [26] created H-SWN (Hindi-SentiWordNet) using two lexical resources namely English SentiWordNet and English-Hindi WordNet Linking. Using WordNet linking they replaced words in English SentiWordNet with equivalent Hindi words to get H-SWN.

4. Challenges in Database Preparation

The important issues involved in database preparation for development of various speech technologies are (i) creating a generic acoustic database that covers language variations and (ii) designing of the recording prompts and recording of speech databases to be used by corpus-based speech synthesizers. The problem arises in collection and selection of texts related with the application domains, the selection of appropriate speakers, all the necessary techniques such as recording setup for assuring and maintaining the same quality during the multiple recording sessions of the resulting database. The high cost of the recording process limits the ability in creating databases in more than a few voices for each domain specific application.

5. Conclusions

In this paper, a survey of efforts in database developments for Hindi language has been performed. It discusses some core linguistic resources of Hindi language, available through various resources developed for usage in text-to-speech synthesis and speech recognition technology. Despite the fact that there are tremendous challenges for building resources according to the global standards, there is immense potential for the development of language resources and technologies in India. If one needs to record his own database, he needs recording equipment (the higher quality, the better). A proper recording studio is ideal, though may not be available for everyone. A cheap microphone stuck on the back of a standard PC is not ideal. A high-quality sound board, close-talking and high-quality microphone and a nearly soundproof recording environment will often be the compromise between these two extremes. The high cost of recording process limits the ability of the technology providers to produce more than a few voices in a particular language. A solution to this problem has been conducted to separate the speech synthesizer from the inventory that defines the synthesizers' voices. Therefore, selection of the inventory of the recordings must be designed to provide good coverage of phonetics and phonation of the selected language, using analysis on available text corpora, mainly newspaper text, books etc for unrestricted TTS. For a restricted domain, domain adaptation is done at speech inventory level. By selecting an inventory with carefully-selected sentences of a restricted domain, such as banking, health care, security, travel sector and others, a very high quality can be achieved for sentences in that domain.

It is suggested that the recordings from various resources can be grouped into application domains that can be combined to generate inventories which can be integrated with speech synthesizers to develop TTS and speech recognition applications. Unfortunately, many of the existing corpora or resources lack features that are strongly desirable for their uses in the scientific context. These shortcomings include problems with availability (in some cases the use of very specific interfaces is required), high costs or strict licenses that permit reuse and data aggregation. This paper identifies the distribution constraints, a challenge for open distribution, which needs to be addressed. As some of these problems can't be removed such as that of copyright, it would be beneficial to have more resources available electronically that can be used with fewer restrictions. This shall enable the participation of a larger group of institutions (within and outside of India) and the industry, as well as in research and development towards building speech systems in Hindi language.

References

[1] Dash, N.S. and Choudhary, B.B. (2011) Why Do We Need to Develop Corpora for Indian Languages? *Proceedings of the International Conferences on SCALLA* (*Vol. 11*), Bangalore.

[2] Kishore, P., *et al.* (2012) The IIIT-H Indic Speech Databases. *Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association*, Portland, 9-13 September 2012, 1-4.

[3] Agrawal, S.S. (2010) Recent Developments in Speech Corpora in Indian languages: Country Report of India. *Proceedings of O-COCOSDA 2010*, Kathmandu, 25 November 2010.

[4] Mukherjee, M. (1996) The First Computer in India. In: Banerjee, U., Ed., *Computer Education in India—Past, Present and Future*, Concept Publications, New Delhi, 13-16.

[5] Sinha, M.K. (2009) A Journey from Indian Scripts Processing to Indian Language Processing. *IEEE Annals of the History of Computing*, **31**, 8-31. https://doi.org/10.1109/MAHC.2009.1

[6] Hindi Universal Word (UW) Dictionary. http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php

[7] Rao, S. (2011) Application Prosody Model for Developing Speech System. *International Journal of Speech Technology*, **11**, 2011.

[8] Quasthoff, U., Mitra, R., Mitra, S., Eckart, T., Goldhahn, D., Goyal, P. and Mukherjee, A. (2012) Large Web Corpora of High Quality for Indian Languages. Proceedings of the 8th International Conference on Language Resources and Evaluation (LERC), Istanbul, 21-27 May 2012, 47.

[9] Kurian, C. (2015) A Review on Speech Corpus Development for Automatic Speech Recognition in Indian Languages. *International Journal of Advanced Networking and Applications*, **6**, 2556.

[10] Arora, S., Saxena, B., Arora, K. and Agarwal, S.S. (2010) Hindi ASR for Travel Domain. *Proceedings of O-COCOSDA* 2010, Kathmandu, 25 November 2010.

[11] Agrawal, S.S. (2010) Recent Developments in Speech Corpora in Indian Languages: Country Report of India. *Proceedings of O-COCOSDA 2010*, Kathmandu, 25 November 2010.

[12] Linguistic Data Consortium for Indian Languages (LDC-IL). http://www.ldcil.org/resourcesSpeechCorpHindi.aspx

[13] Samudravijay, K., Rao, P.V.S. and Agrawal, S.S. (2000) Hindi Speech Data. *Proceedings of the 6th International Conference on Spoken Language Processing* (*ICSLP*), Beijing, 16-20 October 2000.

[14] Agrawal, S.S., Sinha, S., Singh, P. and Olsen, J. (2012) Development of Text and Speech Database for Hindi and Indian English Specific to Mobile Communication Environment. *Proceedings of the International Conference on the Language Resources and Evaluation Conference* (*LREC*), Istanbul, 21-27 May 2012.

[15] The EMILLE Project (Enabling Minority Language Engineering). http://www.emille.lancs.ac.uk/

[16] Hussain, S. (2008) Resources for Urdu Language Processing. *Proceedings of the 6th Workshop on Asian Language Resources*, Hyderabad, 11-12 January 2008, 99-100.

[17] http://corpora.uni_leipzig.org

[18] www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php

[19] Arora, K., Arora, S., Verma, K. and Agrawal, S.S. Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages. Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP), Jeju Island, 4-8 October 2004, 2885-2888.

[20] Syntax and Morphology in Hindi and Urdu: A Lexical Resource. https://clas.uiowa.edu/linguistics/hindi-verb-project

[21] Shabdkosh. http://www.shabdkosh.com/

[22] http://www.shabdkosh.com/content/category/downloads/

[23] Das, A and Bandyopadhyay, S. (2010) SentiWordNet for Indian Languages. *Proceedings of the 8th Workshop on Asian Language Resources* (*ALR*), Beijing, 21-22 August 2010, 1-8.

[24] Hindi Wordnet. http://www.cfilt.iitb.ac.in/wordnet/webhwn/API_downloaderInfo.php

[25] Jha, S., Narayan, D., Pande, P. and Bhattacharyya, P.A. (2001) WordNet for Hindi. Proceedings of the International Workshop on Lexical Resources in Natural Language Processing, Hyderabad, January 2001.

[26] Joshi, A., Balamurali, A.R. and Bhattacharyya, P. (2010) A Fall-Back Strategy for Sentiment Analysis in Hindi: A Case Study. *Proceedings of the Fifth International Conference on Systems* (*ICONS*), Menuires, 11-16 April 2010, 1-6.