# Resources for Indian languages

Arun Baby, Anju Leela Thomas, Nishanthi N L, and TTS Consortium**

Computer Science and Engineering
IIT Madras
arunbaby@cse.iitm.ac.in
{anjuthomas95,nlnishanthi}@gmail.com

**Abstract.** This paper discusses a consortium effort with the design of database for a high-quality corpus, primarily for building text to speech(TTS) synthesis systems for 13 major Indian languages. Importance of language corpora is recognized since long before in many countries. The amount of work in speech domain for Indian languages is comparatively lower than that of other languages. This demands the speech corpus for Indian languages. The corpus presented here is a database of speech audio files and corresponding text transcriptions. Various criteria are addressed while building the database for these languages namely, optimal text selection, speaker selection, pronunciation variation, recording specification, text correction for handling out-of-the-vocabulary words and so on. Furthermore, various characteristics that affect speech synthesis quality like encoding, sampling rate, channel, etc is considered so that the collected data will be of high quality with defined standards. Database and text to speech synthesizers are built for all the 13 languages, namely, Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Odiya, Rajasthani, Tamil and Telugu.

**Key words:** Indian languages, Database, Hybrid segmentation, pruning, CLS, HTS, TTS

## 1 Introduction

Speech is the most prominent means of communication between humans and has the potential of being used as an interface with computers as a mode of interaction [17]. Human beings have always been fancied by the idea to create systems that can understand and talk like humans. Since 1960's, researchers have been trying to develop systems that can interpret and understand human speech. Speech technology plays a crucial role in the development of applications

** *TTS Consortium*: Dr. Hema A Murthy, IIT Madras; Dr. Kishore S Prahallad, IIIT Hyderabad; Dr. K Sreenivasa Rao, IIT Kharagpur; Dr. Chandra Sekhar Seelamantula, IISc Bangalore; Shri. Bira Chandra Singh, CDAC Mumbai; Shri. V K Bhadran, CDAC Thiruvananthapuram; Dr. S R M Prasanna, IIT Guwahati; Shri. Arup Saha, CDAC Kolkata; Dr. T Nagarajan, SSNCE Chennai; Dr. Hemant Patil, DA-IICT Gujarat; Dr. Anil Kumar Sao, IIT Mandi; Dr. V Ramasubramanian, PESIT Bangalore

in various domains like agriculture, health care and government services for common people in a multilingual society like India which has about 22 major languages, written in 13 different scripts, with over 1600 languages/dialects. 13 Indian languages are choosen based on the number of speakers. Much has been discussed on Indian languages [19], its language families [20], nature of scripts, common phones in [5] [2] [1].

Corpus is the machine readable form of the large collection of structured text in written or spoken form [4]. The importance of language corpora is recognized since long before in many countries. The amount of work in speech domain for Indian languages is comparatively lower than that of other languages. A corpus for Indian languages is a time taking process because of lack of resources and is difficult because of its diversity. However, there exists a lot of scope in developing language corpora for Indian languages. The information acquired from the corpora will not only provide advanced resources for developing language processing tools, TTS, etc. but also will be useful for education purposes and various domains of language research.

An initiative was taken by DeiTY, Ministry of Information Technology, India to sponsor the development of TTS in regional languages and as a part of which data collection is carried out. The speech data for the database is collected by the joint effort of all the consortium members. The consortium members include IIT Madras, IIIT Hyderabad, IIT Kharagpur, IISc Bangalore, CDAC Mumbai, CDAC Thiruvananthapuram, IIT Guwahati, CDAC Kolkata, SSNCE Chennai, DA-IICT Gujarat, IIT Mandi and PESIT Bangalore. For speech recording, two voice talents are identified(1 male and 1 female) for each language. Text in each language is identified for reading and is read in an anechoic chamber. A total of 40 hours of speech data is collected for a language - 20 hours of native(mono) data (10 hours each of male and female data) and 20 hours of English data recorded by these native speakers (10 hours each of male and female data).

This paper describes the development of speech corpora/database for 13 Indian languages by the TTS Consortium. It also describes the tools, developed at IIT Madras, that aid the process of building TTS systems. The various issues while developing speech corpora like the selection of text, text processing, the purpose of use, selection of speakers, the manner of data collection, the size of corpora, issues in transcription, type of data encoding, the manner of data-sampling, etc are mentioned in Section 2. Section 3 gives an overview of building TTS Systems. Section 4 details the Android applications built for TTS. Section 5 concludes the work.

## 2 Data Collection

This section discusses the design and methodology used for the collection of the speech database. It also describes the procedures for text collection, text correction(if any), the methodology followed for selecting the speakers and details of the recording process.

### 2.1 Text selection/correction

A huge text corpus is a very crucial resource for preparing the training data for building TTS. Collecting transcribed data for Indian languages, which are in general low resource languages, is a herculean task. To accomplish this task, initially, text in various Indian languages are collected from newspapers, websites, blogs, etc with the help of web crawlers. Furthermore, text from different domains like children stories, literature, science, tourism, etc was also manually collected to achieve a good coverage. The collected text is manually corrected to get rid of transcription errors if any. Care has also been taken to ensure that the chosen text is easy to read, covers the most commonly used words and phrases in a language and has maximum syllable coverage. The collected data is used to record speech data for speech synthesis system building processes and also to generate a list of words for pronunciation dictionary.

### 2.2 Speaker selection

The next phase, after text collection/selection, is to record the data. 2 voice talents(1 male and 1 female speaker) are selected, for each of the regional languages, in such a way that there are minimal pronunciation errors. Multi-speaker recording for given language, gender and type(mono/English) will lead to variations in pitch, speed, speech style, tempo and amplitude. Single speaker data limits the variations and change in voice quality. Apart from these aspects, appropriate voice talent whose voice seems pleasant to listen, as well as amenable to signal processing is chosen. For the ease of speaker, care has been taken in every session to keep the context of the text co-relative(unchanged) so that switching is avoided.

### 2.3 Recording

The recording is carried out in a special environment which is free from noise and echo. A speech corpus is collected for 13 languages, each containing 40 hours(native and English of Male and Female speakers respectively) of data. The recording is done by professional speakers(male and female) to maintain constant pitch and prevent stress phenomenon. Further, to avoid the fatigue of the speaker, a break is given every 45 minutes. Later the recorded sentences are split at the sentence level. Also, measures have been taken to maintain same conditions and voice characteristics across the multiple recording sessions. Hence, the type of recording is mono, with a sampling rate of 48KHz and the number of significant bits per sample is 16. Non-conforming sentences could be re-recorded. The recorded speech files are stored in .wav format in the database.

### 2.4 Summary of the text corpus

Table 1 gives a summary of the data collected in hours, the number of words and number of sentences for each language.

**Table 1.** Summary of the corpus

| | | **Female** | | **Male** | |
|---|---|---|---|---|---|
| **Languages** | | *English* | *Mono* | *English* | *Mono* |
| **Assamese** | Duration in hours | 12.05 | 14.45 | 11.30 | 12.95 |
| | Number of words | 17531 | 29510 | 18143 | 32136 |
| | Number of sentences | 8513 | 8713 | 8892 | 8941 |
| **Bengali** | Duration in hours | 5.2 | 5.01 | 10.03 | 10.05 |
| | Number of words | 8607 | 18599 | 12901 | 30493 |
| | Number of sentences | 3239 | 3253 | 5316 | 6187 |
| **Bodo** | Duration in hours | - | 4 | - | - |
| | Number of words | - | 3991 | - | - |
| | Number of sentences | - | 2715 | - | - |
| **Gujarati** | Duration in hours | 10 | 10.33 | 10.13 | 10.92 |
| | Number of words | 14309 | 20567 | 15192 | 23546 |
| | Number of sentences | 4671 | 2396 | 4826 | 3288 |
| **Hindi** | Duration in hours | 7.94 | 7.23 | 7.81 | 7.03 |
| | Number of words | 15153 | 13380 | 15189 | 13369 |
| | Number of sentences | 5240 | 2605 | 5243 | 2806 |
| **Kannada** | Duration in hours | 7.5 | 11.82 | 7.48 | 7.03 |
| | Number of words | 13738 | 11097 | 14446 | 11358 |
| | Number of sentences | 4448 | 5132 | 4778 | 5934 |
| **Malayalam** | Duration in hours | 8.77 | 8.19 | 7.89 | 9.7 |
| | Number of words | 13738 | 29165 | 13738 | 28933 |
| | Number of sentences | 5132 | 5650 | 5131 | 5650 |
| **Manipuri** | Duration in hours | 10.35 | 10.14 | 10.22 | 10.61 |
| | Number of words | 21119 | 23555 | 18535 | 24531 |
| | Number of sentences | 10167 | 9487 | 9836 | 9745 |
| **Marathi** | Duration in hours | - | 4.8 | - | 3.27 |
| | Number of words | - | 18287 | - | 12201 |
| | Number of sentences | - | 2448 | - | 1992 |
| **Odia** | Duration in hours | - | 4.27 | - | 4.47 |
| | Number of words | - | 3936 | - | 4069 |
| | Number of sentences | - | 3578 | - | 3573 |
| **Rajasthani** | Duration in hours | 7.25 | 10.24 | 7.30 | 9.82 |
| | Number of words | 11929 | 20923 | 13114 | 22894 |
| | Number of sentences | 3830 | 4346 | 4809 | 4779 |
| **Tamil** | Duration in hours | 12.7 | 10.03 | 10.9 | 10.3 |
| | Number of words | 20911 | 28817 | 20220 | 32017 |
| | Number of sentences | 7914 | 3243 | 7547 | 3717 |
| **Telugu** | Duration in hours | - | 23.92 | - | 4.2 |
| | Number of words | - | 42063 | - | 12192 |
| | Number of sentences | - | 4043 | - | 2481 |

# 3 Voice building

## 3.1 CLS

A common label set (CLS), that capitalizes on the acoustic similarity of Indian languages, is devised using the Latin-1 script. The CLS provides a standardized representation for phonemes across different Indian languages. Phones that are similar are mapped to the same label. The notations and logic used in deriving the CLS are detailed in [15] [8].

## 3.2 Parsing and Unified parser

The traditional parsing approach uses the respective language's rules to parse the word into corresponding phones. This uses a sequential left-to-right approach in parsing the words. This approach has some limitations. A unified approach which uses the generic language structure of Indian languages is developed [3] [12]. The unified parser attempts to unify the languages based on the Common Label Set. It is designed across all the languages capitalizing on the syllable structure of Indian languages. The Unified Parser converts UTF-8 text to common label set, applies letter-to-sound rules and generates the corresponding phoneme sequences. Given the unity in the diversity of Indian languages, developing parsers for new languages is easy using the unified approach.

## 3.3 Hybrid segmentation

Accurate realization of phoneme segments plays a key role in speech synthesis systems, as this information is used in duration modeling in Hidden Markov Models (HMMs). For low-resource languages, the accuracy is generally ensured through manual labeling. Manual correction is a monotonous task. When labeled by different people, it yields inconsistent output too. Automatic segmentation is introduced to overcome these issues. Flat-start initialization of monophone HMMs, Embedded reestimation and Forced-Viterbi alignment are the three steps used in conventional segmentation. But this model does not indicate the boundary positions. Syllables are the fundamental units of speech production and perception. Hence, the syllable boundaries will be more distinct than that of phonemes. The acoustic energy is significantly low at syllable boundaries. This enables the use of short term energy (STE) as a measure to determine the syllable boundaries. Nevertheless, local fluctuations do not allow the direct application of STE in detecting the boundaries precisely. Therefore group delay processing is used to smooth the STE function, after which, this is used to detect the syllable boundaries. The boundaries of the syllables are corrected with group delay and spectral flux. Syllable splicing, reestimation of models within syllables and syllable-level forced Viterbi alignment is done after the boundary correction. A two-pass procedure is followed to perform this boundary correction and reestimation. The process of hybrid segmentation is detailed in [16] [9].

### 3.4 Pruning

Pruning is the process of discarding badly segmented units from a database using the acoustic properties of syllable units. Duration, average f0 and STE are the cues taken into consideration [14]. Pruning helps in the correction of segmentation errors and also in maintaining acoustic continuity in the database.

The use of pruned units to initialise HMMs results in a considerable improvement in the quality of speech synthesizers, as only the well-articulated phonemes take part in the synthesis [11].

### 3.5 HTS

HMM-based speech synthesis system (HTS) is a statistical parametric approach [18]. It involves obtaining a parametric representation of speech by extracting the spectral and excitation features from the database. To synthesize a given text, instead of concatenating pre-recorded speech units, the speech waveform is derived from the parametric representations of speech. The voices built are hosted here [13].

## 4 Android Applications

Three Android applications were developed to make the TTS services available in Android platform : (1) Tamil TTS app - for Tamil text-to-speech synthesis (2) Hindi TTS app (for Hindi text-to-speech synthesis) and (3) Indic TTS app - for text-to-speech synthesis of 13 Indian languages mentioned in the previous sections. (1) and (2) were developed initially with the use language-specific parsers. The development of unified parser gave way to the idea of a single Android app, that can handle all the 13 languages, instead of language-specific apps. (1) and (2) takes Tamil and Hindi text, respectively, as input and reads it out to the user. (3) takes input text in any of the 13 Indian languages and reads it out to the user. These apps can be used as stand-alone applications or can be used in conjunction with mobile browsers, messengers,etc. The apps are available for download in the Indic TTS website [7].

## 5 Conclusion

Speech corpus applications have tremendous prospects in India. An attempt has been made, through this paper, to give a comprehensive survey of the development of a corpus for speech synthesis in Indian languages. The data is hosted on the web [6], under the terms and conditions mentioned in [10], hoping that these resources would be available to all groups of people working for corpus generation and research activities.

## Acknowledgment

We would like to thank the Department of Information Technology, Ministry of Communication and Technology, Government of India for funding the project Development of Text-to-Speech Synthesis for Indian Languages PhaseII (Ref. no. 11(7)/2011-HCC(TDIL)), as part of which, most of the research techniques have been developed.

We also thank and appreciate the contributors - Dr. Hema A Murthy, IIT Madras for Hindi and Bodo; Dr. Kishore S Prahallad, IIIT Hyderabad for Telugu data; Shri. Bira Chandra Singh, CDAC Mumbai for Marathi and Odia data; Shri. V K Bhadran, CDAC Thiruvananthapuram for Malayalam data; Dr. S R M Prasanna, IIT Guwahati for Manipuri and Assamese data; Shri. Arup Saha, CDAC Kolkata for Bengali data; Dr. T Nagarajan, SSNCE Chennai for Tamil data; Dr. Hemant Patil, DA-IICT Gujarat for Gujarathi data; Dr. Anil Kumar Sao, IIT Mandi for Rajasthani data; Dr. V Ramasubramanian, PESIT Bangalore for Kannada data - for their leadership and contribution towards building the Indian language corpora.

## References

1. SS Agrawal, Sunita Arora, and Karunesh Arora. Towards design, development and standardization of speech corpora for developing indian language tts system. *COCOSDA-2005, Dec*, pages 6–8, 2005.
2. Karunesh Arora, Sunita Arora, Kapil Verma, and Shyam Sunder Agrawal. Automatic extraction of phonetically rich sentences from large text corpus of indian languages. In *INTERSPEECH*. ISCA, 2004.
3. Arun Baby, Nishanthi N L, Anju Leela Thomas, and Hema A Murthy. A unified parser for developing indian language text to speech synthesizers. In *International Conference on Text, Speech and Dialogue*. Springer, 2016.
4. NILADRI SEKHAR Dash and BIDYUT BARAN Chaudhuri. Why do we need to develop corpora in indian languages. In *International Conference on SCALLA, Banglore*, 2001.
5. Prahallad Lavanya, Prahallad Kishore, and Ganapa Thiraju Madhavi. A simple approach for building transliteration editors for indian languages. *Journal of Zhejiang University Science A*, 6(11):1354–1361, 2005.
6. IIT Madras. Indic tts. `https://www.iitm.ac.in/donlab/tts/`.
7. IIT Madras. Indic tts - android apps. `https://www.iitm.ac.in/donlab/tts/androidapp.php`.
8. IIT Madras. Indic tts - cls. `https://www.iitm.ac.in/donlab/tts/cls.php`.
9. IIT Madras. Indic tts - hybrid segmentation. `https://www.iitm.ac.in/donlab/tts/hybridSeg.php`.
10. IIT Madras. Indic tts - license. `http://www.iitm.ac.in/donlab/tts/downloads/license.pdf`.
11. IIT Madras. Indic tts - pruning. `https://www.iitm.ac.in/donlab/tts/prune.php`.
12. IIT Madras. Indic tts - unified parser. `https://www.iitm.ac.in/donlab/tts/unified.php`.

13. IIT Madras. Indic tts - voices. https://www.iitm.ac.in/donlab/tts/voices.php.
14. K Raghava Krishnan. Prosodic analysis of Indian languages and its application to text to speech synthesis. http://lantana.tenet.res.in/thesis.php, M S Thesis, Department of Electrical Engineering, IIT Madras, India, July 2015.
15. B Ramani, S Lilly Christina, G Anushiya Rachel, V Sherlin Solomi, Mahesh Kumar Nandwana, Anusha Prakash, S Aswin Shanmugam, Raghava Krishnan, S Kishore, K Samudravijaya, et al. A common attribute based unified hts framework for speech synthesis in indian languages. In *8th ISCA Workshop on Speech Synthesis*, pages 311–316, 2013.
16. S Aswin Shanmugam and Hema Murthy. A hybrid approach to segmentation of speech using group delay processing and hmm based embedded reestimation. *presentation in INTERSPEECH*, 2014.
17. Pukhraj Shrishrimal, Ratnadeep R Deshmukh, and Vishal Waghmare. Indian language speech database: a review. *International Journal of Computer Application (IJCA)*, 47(5):17–21, 2012.
18. Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1315–1318. IEEE, 2000.
19. Wikipedia. Languages of india. https://en.wikipedia.org/wiki/Languages_of_India.
20. Wikipedia. South asian language families. https://en.wikipedia.org/wiki/Languages_of_South_Asia.