The IIIT-H Indic Speech Databases

Kishore Prahallad[1], *E.Naresh Kumar*[1], *Venkatesh Keri*[1], *S.Rajendran*[1], *Alan W Black*[2]

[1]Speech and Vision Lab, International Institute of Information Technology, Hyderabad, India
[2] Language Technologies Institute, Carnegie Mellon University, Pittsburg, USA

kishore@iiit.ac.in, nareshkumar.elluru@research.iiit.ac.in, venkateshk@research.iiit.ac.in
su.rajendran@gmail.com, awb@cs.cmu.edu

Abstract

This paper discusses the efforts in collecting speech databases for Indian languages – Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu. We discuss relevant design considerations in collecting these databases, and demonstrate their usage in speech synthesis. By releasing these speech databases in the public domain without any restrictions for non commercial and commercial purposes, we hope to promote research and developmental activities in building speech synthesis systems in Indian languages.

Index Terms:speech databases, speech synthesis, Indian languages

1. Introduction

Twenty two languages have an official status in India. Apart from these 22 official languages, there also exists several hundred languages and dialects. A few of these languages are spoken by millions. In such a large multilingual society, speech and language technologies play an important role in enabling information access to the illiterate using text-to-speech conversion, and in information exchange using speech-to-speech translation systems. Efforts are on 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 [1]. 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.

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. This enables participation of a larger group of institutions (within and outside of India) and the industry, in research and development towards building speech systems in Indian languages. A common set of speech databases act as benchmark speech databases to compare, evaluate and share knowledge across the institutions. To our knowledge, there has been no such works or efforts in the past in the context of Indian languages.

As of now, we have developed speech databases for Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu. In this paper, we discuss the design issues involved in the development of these speech databases. We show their application in building speech synthesis systems and highlight issues and problems that could be further addressed.

2. Scripts and sounds of Indian languages

The scripts for Indian languages have originated from the ancient Brahmi script. The basic units of the writing system are referred to as Aksharas. The properties of Aksharas are as follows: (1) An Akshara is an orthographic representation of a speech sound in an Indian language; (2) Aksharas are syllabic in nature; (3) The typical forms of an Akshara are V, CV, CCV and CCCV, and thus have a generalized form of C*V. Here C denotes a consonant and V denotes a vowel.

2.1. Convergence and divergence

Most of the languages in India, except (English and Urdu) share a common phonetic base, i.e., they share a common set of speech sounds. This common phonetic base consists of around 50 phones, including 15 vowels and 35 consonants. While all of these languages share a common phonetic base, some of the languages such as Hindi, Marathi and Nepali also share a common script known as Devanagari. But languages such as Telugu, Kannada and Tamil have their own scripts.

The property that separates these languages can be attributed to the phonotactics in each of these languages, rather than the scripts and speech sounds. Phonotactics are permissible combinations of phones that can co-occur in a language. This implies that the distribution of syllables encountered in each language is different. Prosody (duration, intonation, and prominence) associated with a syllable is another property that separates these Indian languages significantly.

2.2. Digital representation

Prior to Unicode, there were several representations for scripts in Indian languages. This included several fonts for each script and several mechanisms (soft keyboards, keyboard layouts and transliteration schemes) of keying the script using QWERTY keyboard [2]. With the advent of Unicode, the scripts of Indian languages have their own unique representation. This has standardized the representation of Aksharas and their rendering on the computer screen.

However, the key-in mechanism of these Aksharas has not been standardized. It is hard to remember and key-in the Unicode of these scripts directly by a layman user of a computer. Thus, soft keyboards, keyboard layouts on top of QWERTY keyboards are still followed. Transliteration scheme, i.e., mapping the Aksharas in Indian languages to English alphabets to key-in is another popular mode. Once these Aksharas are keyed-in, they are internally processed and converted into Unicode characters. Due to this non-standardization, the key-in mechanism of Indian language scripts has to be addressed explicitly during the development of text processing modules in text-to-speech systems and user interfaces.

3. Development of speech databases

The following are the design choices we made in development of these speech databases.

- **Public domain text:** Most of the texts available in Indian languages are in the form of News data or blogs which are under copyright. Hence, we choose to use Wikipedia articles in Indian languages as our text corpus. The articles of Wikipedia are in the public domain. We could select a set of sentences, record speech data and release in public domain without any copyright infringements.
- **Choice of language and dialect:** We used Wikipedia dump of Indian languages released in 2008. This dump consists of 17 Indian languages. We chose to build speech database for Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu. These languages were chosen, as the total number of articles in each of these languages were more than 10,000 and native speakers of these languages were available in the campus. Table 1 shows the statistics of text corpus collected for these languages.
- **Speaker selection:** To record the speech database, a process of speaker selection was carried out. A group of four to five native speakers (who volunteered for speech data collection) was asked to record 5-10 minutes of speech. A speaker was selected based on how pleasant the voice was and how amenable the speech was for signal processing manipulations.

  Each of these languages have several dialects. As a first step, we chose to record the speech in a dialect the native speaker was comfortable with. The native speakers who volunteered to record speech data were all in the age group of 20-30. During the recording process, they were made aware that the speech data being recorded would be released in public domain and a written consent was taken.

3.1. Optimal Text Selection

Given the text corpus in each language, a set of 1000 phonetically balanced sentences were selected as described in [3]. This optimal set was selected using Festvox script that applies the following criteria.

- Each utterance should consists of 5-15 words.
- Each word in the utterance should be among the 5000 most frequent words in the text collection.
- No strange punctuation, capitals at the beginning, and punctuations at the end.

Table 2 shows the statistics of optimal text selected for each of the languages.

3.2. Speech Recording

The speech data was recorded in a professional recording studio using a standard headset microphone connected to a Zoom handy recorder. We used a handy recorder as it was highly mobile and easy to operate. By using a headset the distance from the microphone to a mouth and recording level was kept constant.

A set of 50 utterances were recorded in a single wave file. After each utterance, the speaker was instructed to pause briefly and start the next utterance. This avoided the start-stop for each utterance. The recording was typically clean and had minimal background disturbance. In spite of care being taken, there were mistakes in the utterances due to wrong pronunciation or repeated pronunciation of a word. Any mistakes made while recording were rectified either by re-recording those utterances or by correcting the corresponding transcription to suit to the utterance.

3.2.1. Audio file segmentation

As each wave file consisted of at least 50 utterances, we used the zero frequency filtering (ZFF) technique to automatically segment into utterances. ZFF has been shown to detect voiced and unvoiced regions in a speech signal with high accuracy [4]. The duration of unvoiced regions was subjected to a threshold. This resulted in slicing each

Table 1: *Statistics of the Wikipedia text corpus.*

Languages	No.of sentences	No.of words		No.of Syllables		No.of Phones	
		Total	Unique	Total	Unique	Total	Unique
Bengali	54825	1830902	510197	1689005	4883	2851838	47
Hindi	44100	1361878	376465	942079	6901	1466610	58
Kannada	30330	360560	257782	3037748	5580	1697888	52
Malayalam	84000	1608333	699390	3157561	15259	5352120	51
Marathi	30850	810152	270913	1012066	2352	1452175	57
Tamil	99650	1888462	857850	3193292	10525	5688710	35
Telugu	90400	2297183	763470	3193292	9417	4940154	51

Table 2: *Statistics of the optimal text selection.*

Languages	No.of sentences	No.of words		No.of Syllables		No.of Phones		Avg.Words per line
		Total	Unique	Total	Unique	Total	Unique	
Bengali	1000	7877	2285	25757	866	37287	47	7
Hindi	1000	8273	2145	19771	890	30723	58	8
Kannada	1000	6652	2125	25004	851	37651	51	6
Malayalam	1000	6356	2077	21620	1191	38548	48	6
Marathi	1000	7601	2097	25558	660	37629	57	7
Tamil	1000	7045	2182	23284	930	42134	35	7
Telugu	1000	7347	2310	24743	997	40384	51	7

Table 3: *Duration of speech databases.*

Language	Duration (hh:mm)	Avg. Dur of each utterance(sec)
Bengali	1:39	5.94
Hindi	1:12	4.35
Kannada	1:41	6.05
Malayalam	1:37	5.83
Marathi	1:56	6.98
Tamil	1:28	5.27
Telugu	1:31	5.47

wave file into 50 utterances. A manual check was followed to ensure that each of the utterances match with the corresponding text.

Table 3 shows the total duration of speech database and average duration of each utterance for all the languages.

4. Building synthetic voices

The issues involved in building synthetic voices for Indian languages are as follows – 1) definition of phone set and acoustic-phonetic properties of each phone, 2) letter-to-sound rules, 3) syllabification rules, 4) prominence marking of each syllable, 5) phrase break prediction, 6) choice of unit size in synthesis and 7) prosody modeling. While there is some clarity on the phone set and corresponding acoustic-phonetic feature, rest of the issues are largely unexplored for speech synthesis in Indian languages.

To build prototype voices, we used IT3 transliteration scheme to represent the scripts of Indian languages. A phone set was defined for each language based on our experience. Table 4 shows the acoustic-phonetic features defined for these phones.

The concept of letter-to-sound (grapheme-to-phoneme or Akshara-to-sound) rules is more applicable to Bengali, Hindi, Malayalam, Marathi, Tamil than to Kannada and Telugu. Moreover, there hardly exists a decent set of letter-to-sound rules that one could use readily. Hence, we did not use any letter-to-sound rules in this phase of building voices. Our hope was that phone level units when clustered based on the context, would produce appropriate sound.

Syllabification is another issue. One could use Akshara as an approximation of syllable. It is known that acoustic syllables differ from Aksharas. For example the Aksharas of the word /amma/ (meaning mother) correspond to /a/ /mma/. However, acoustic syllables are /am/ and /ma/. Given that syllabification is specific to each language, we used Aksharas as syllables in these current builds.

Indian languages are syllable-timed, as opposed to stress-timed languages such as English. Hence, the concept of syllable-level prominence is more relevant for Indian languages. Prominence pattern plays an important

Table 4: *Acoustic phonetic features.*

Feat. Name	Feat. Values	Range
Phone type	Vowel/Consonant	2
Vowel length	Short/Long/dipthong/schwa	4
Vowel height	High/middle/low	3
Vowel frontness	Front/mid/back	3
Lip rounding	+/-	2
Consonant type	Stop/fricative/affricatives/ nasal/lateral	5
Place of articulation	Labial/alveolar/palatal/ labio-dental/dental/velar	6
Consonant voicing	voiced/unvoiced	2
Aspiration	+/-	2
Cluster	+/-	2
Nukta	+/-	2

role in text-to-speech systems. Given that there is hardly any research on syllable-level prominence for Indian languages, we assigned primary prominence to first syllable in the word. Rest of syllables were assigned second prominence.

Prediction of breaks aids intelligibility and naturalness of synthetic voices. It should be noted that prosodic phrase breaks differ significantly from syntactic phrase breaks. An appropriate modeling of prosodic phrase breaks requires part-to-speech (POS) tags. A POS tagger is hardly available for all Indian languages. In the current build, we have used punctuation marks as indicators of phrase breaks.

Given that basic units of writing systems in Indian languages are syllable-like units (Aksharas), the choice of syllable versus phone needs to be investigated further in detail, for statistical parametric synthesis. In the current build, we used phone as a unit for both the unit selection and statistical parametric voices. Prosody modeling for expressive style synthesis is also an important issue to be addressed.

With the choices made, we built unit selection (CLUNITS) and CLUSTERGEN voices for these Indian languages in Festvox framework [5][6]. Table 5 shows the objective evaluation of these voices in terms of Mel-cepstral distortion (MCD). CLUSTERGEN voices have lower MCD scores than CLUNITS. This is primarily because of use of natural durations in MCD computation for CLUSTERGEN voices. Among the CLUSTERGEN voices, Hindi and Tamil voices have higher MCD scores. This could be attributed to lack of appropriate letter-to-sound rules in these builds. However, it was interesting to note a lower MCD score for CLUSTERGEN voice of Bengali, in spite of not using any letter-to-sound rules. The voices of Telugu and Marathi have the best MCD, but both are affected by excessive silence in the recording (between words), and have a lower intelligibility in comparison with other voices.

Table 5: *Mel-cepstral distortion (MCD) scores for CLUNITS and CLUSTERGEN voices.*

Languages	MCD	
	clunits	cg
Bengali	7.74	4.96
Hindi	7.09	5.24
Kannada	6.90	5.01
Malayalam	7.78	5.1
Marathi	7.08	4.4
Tamil	8.0	5.30
Telugu	6.55	4.39

5. Conclusion

We have discussed the design choices made in development of speech databases for seven Indian languages. Also, we have highlighted a set of research issues or topics that could be addressed in the context of building speech synthesis systems for these languages. A set of baseline voices were also built to demonstrate the feasibility. These voices, text, and speech databases are available for download from http://speech.iiit.ac.in and http://festvox.org, under the public domain. We hope that the release of these databases will be helpful for the speech community within India and abroad towards the development of speech systems in Indian languages.

6. Acknowledgements

This work was partly funded and supported by Microsoft Research Young Faculty award to Kishore Prahallad at IIIT-H. We would like to thank Prof. Rajeev Sangal (IIIT-H), Prof. B. Yegnanarayana (IIIT-H) and Prof. Hema A. Murthy (IIT Madras) for the encouragement and useful suggestions on this work. We also would like to thank E. Veera Raghavendra for useful discussions on this topic.

7. References

[1] Hema A. Murthy *et. al.*, "Building Unit Selection Speech Synthesis in Indian Languages: An Initiative by an Indian Consortium", in Proceedings of COCOSDA, Kathmandu, Nepal, 2010.

[2] Anand Arokia Raj, Tanuja Sarkar, Satish Chandra Pammi, Santosh Yuvaraj, Mohit Bansal, Kishore Prahallad, Alan W. Black "Text Processing for Text to Speech Systems in Indian Languages", in Proceedings of 6th ISCA Speech Synthesis Workshop SSW6, Bonn, Germany, 2007.

[3] Alan W. Black and Kevin Lenzo, "Building Voices in Festival Speech Synthesis System", `http://festvox.org/bsv/`.

[4] Sri Rama Murty K, B. Yegnanarayana, Anand Joseph Xavier M, "Characterization of Glottal Activity From Speech Signals", IEEE Signal Processing Letters, vol. 16, no. 8, pp. 469-472, June 2009.

[5] Alan W. Black and Paul Taylor "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", in Proceedings of Eurospeech, Rhodes, Greece, 1997.

[6] Alan W. Black "CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modelling", in Proceedings of Interspeech, Pittsburgh, pp. 1762-1765, 2006.