# 8.13 Speech Corpora Development in Indian Languages

*Shyamal Kr. Das Mandal and Arup Saha*
*Centre for Development of Advanced Computing (C-DAC), Kolkata*

***Abstract*—Facts Database + Referential Rules→ Artificial Intelligence (AI). Hence, a collection of standard sample data (or corpora) is crucial for any AI. Similarly, for the development of speech technology and speech research the standard annotated speech database (corpora) plays such an anchoring role. Speech corpora provide the important annotated voice segments to the researcher for making test bed, analysis, and collection of statistics on different speech parameters. Quality Speech Corpora should contain most of the basic elements of speech research like acoustic phonetics and acoustic prosodic; and also the reference data for speech technology development. This paper discusses an approach for creation and management of such Annotated Speech Corpora covering the common needs of speech research and speech technology development including speech recognition, speaker recognition and speech synthesis.**

***Index Terms*—Speech Corpora, Annotation, Informant, Prosody, bi-phones, Suprasegmental.**

## I. INTRODUCTION

SPOKEN language interface to computers is a topic that has lured and fascinated engineers and speech scientists for more than five decades. For many, the ability to converse freely with a machine represents the ultimate challenge to Artificial Intelligence, as the production and perception of human speech constitute the highest form of human cognition. In addition to being a challenging topic, spoken language interfaces are fast becoming a necessity. In the current Indian context, the needed skills of both computer literacy as well as conversance with written English in the standard QWERTY English keyboard-oriented interfaces to Personal Computer (PC), restrict the usage of such essential IT access device to only a miniscule fraction of Indian population. Since speech is a primary mode of direct communication amongst human beings, it would be highly desirable from added convenience, as well as empowerment, points of view to facilitate for such direct speech-anchored dialogue between man and machine. Added empowerment by such speech enablement could be the direct local language support in place of English. This demands full-scale research and development in the area of spoken language processing (including speech recognition and speech synthesis) in different Indian Languages. The development of such Speech Research and Technology on Acoustics/Phonetics of different spoken Indian Languages demands large speech databases for each. The creation of annotated Speech Corpora becomes, therefore, a prime enabler for such research. In case of European and some of the Asian languages the annotated speech corpora and creation methodology are already available in European Lang Resources Assoc [1] and Linguistic Data Consortium (LDC) [2]. In case of English, TIMIT is one of the speech data resources that is used by most of the speech researchers.

While corpora of written text exist in most of the Indian Languages by now, unfortunately none of these exist for their spoken versions. The absence of standard speech corpora in Indian Languages has been one of the major retarding factors slowing down the developments of speech research, particularly related to man-machine-man interactions. In addition, such standard speech corpora can also provide a commonly accessible base line to evaluate the efficacies of current state-of-the-art performance in speech research areas [3].

The major issues for building up such a Speech Corpora are the selection of dialect, content, informants and recording environment [4]. Proper annotation and marking of linguistic and phonetic units are also essential. In the selected dialect, such a speech corpora should adequately cover speaking pattern of native informants, normal as well as under different emotional stresses, for different types of sentences under a variety of discourses.

Recently C-DAC, Kolkata has developed basic speech corpora in Bangla, Assamese and Manipuri Language, which is considered only the essential requirement to support for the technology research and development for man-machine-man communication in Indian Languages. The developed Corpora may appear to be adequate only in this limited sense. Primary restriction is towards the manner of speaking (i.e. Text Reading Mode). The database for the study of supra-segmental characteristics may be inadequate for deriving rules for the purpose of generation of fully natural speech. It may be extremely difficult to mimic spontaneous human speech mechanically. Spontaneous interactions between speakers are not yet included. In this regards C-DAC Kolkata also published a technical report [5][6] for creating such speech corpora. This paper describes those basic procedures and the required augmentation for possible extension of the corpora to support Spoken Language Processing.

## II. PROCEDURE

Development of speech corpora involves essentially four steps, namely: (a) Selection of content or Text to be spoken by a number of informants, (b) Recording Speech in Digital Format, (c) Annotation of collected Speech Data and (d) User-friendly management of the recorded data along with the associated text.

*Shyamal Kr. Das Mandal and Arup Saha are with the Centre for Development of Advanced Computing (C-DAC), Kolkata*
*(e-mail: shyamal.dasmandal@kolkatacdac.in)*

### A. *Selection of Content or Text*

There exists adequate experience in text corpora building in most of the major Indian Languages. This is primarily useful for Natural Language Processing (NLP), the backbone of text-oriented technology development, such as Machine Translation (MT), Searching and Information Retrieval, Automatic Summarizations, etc. Unfortunately Spoken Language Processing (SLP) has an additional dimension that cannot be addressed purely by mimicking these well-understood NLP procedures; here, one must respect the importance of prosodic information present in spoken language and also the gaps/absence of strict grammatical structures of written text. While selecting the contents of Speech Corpora, it is, therefore, essential to adequately address various contrasting differences amongst formalisms of grapheme text vis-à-vis their absence in speech.

However, due to primary emphasis on Speech Corpora contents to various needs of technology development for supporting man-machine-man communication, instead of general linguistic research, the content selection naturally gets governed by those specific technological requirements. In this context, Speech Corpora contents fall largely into following four groups:

- Speech Research
- Automatic Speech Recognition
- Speaker Recognition
- Speech Synthesis
- Language Identification

The schematic in Figure1 illustrates how the above diverse needs generally influence the content selection for speech corpora. The rectangular boxes indicate types of targeted contents, whereas, the interconnecting arrowed lines highlight the intricate interactions supporting different technological vis-à-vis research needs.

### B. *Recording Speech in Digital Format*

The above selected content has to be recorded from different informants in digital domain. The environmental noise plays an important role in recording. On the other hand, the telephone channel voice is bandwidth limited by the channel characteristics. Hence, the recorded samples must contain characteristic noises impregnated within the voice sample. However, to begin with, one can also include normal studio quality recordings.

In case of Bangla Speech Corpora, speech is recorded in studio environment directly in digital format using the shure-make dynamic microphone connected to a PC add-on sound card driven by commonly available recording software in 16-bit PCM mono mode with a sampling frequency 22,050 Hz.

In case of Bangla corpora, recordings had been made for different age groups of both sexes of Bangla Standard Colloquial Bengali (SCB) speaking informant. The selection of the informant had been done based on their test performance of uttering of consonant, vowel, and nasal nominal pair. Adequate web-enabled meta-data of the selected informant is also maintained for later referencing.

### C. *Annotation of Recorded Speech Data*

The term annotation covers any descriptive or analytic notations applied over raw speech data. The added annotations may include various kinds of both acoustic and/or linguistic information; namely, (a) phoneme, (b) syllable, and (c) word boundaries along with the appropriate (d) Parts of Speech (POS) and/or (e) phrase/clause markers in the sound spectrum. Usually, there lacks unanimity over the definitions of phoneme-boundaries in any continuous speech. However, in the present endeavor, the following phoneme marking conventions have been generally followed:

- Consonant phonemes are defined from the beginning of occlusion to the end of Voice-Onset-Time (VOT) for plosives and affricates. Inclusion of VOT here means inclusion of the aperiodic transition, which, inter-alia, is co-articulatory in nature; and hence, may vary for the same consonant in different contexts.
- For nasal consonants, trills, laterals and sibilants are defined as the durations of closure or constriction as the case may be.
- Vowel phonemes are defined as the total vocalic region inclusive of the vocalic transitions.
- For Vowel-to-Vowel (VV) transitions the segment boundary is marked at the middle of the transitory part.
- For diphthongs the segment includes the prominent target and the long transition.

Semi-automatic tagging software is developed for the acoustic tagging which will give a user-friendly environment to the tager [7].

### D. *User-friendly Management of the Recorded Data*

As mentioned earlier, Speech Corpora contents are expected to support four distinct types of investigative research; namely, (a) speech research, (b) automatic speech recognition, (c) speaker recognition, and finally, (d) speech synthesis. Moreover, selected corpora contents of each of the above kinds need to have two interrelated components; namely, text and speech. Such text components are stored as Unicode text and the corresponding speech recordings are stored in PulseCoded Modulation (PCM) in .wav format.

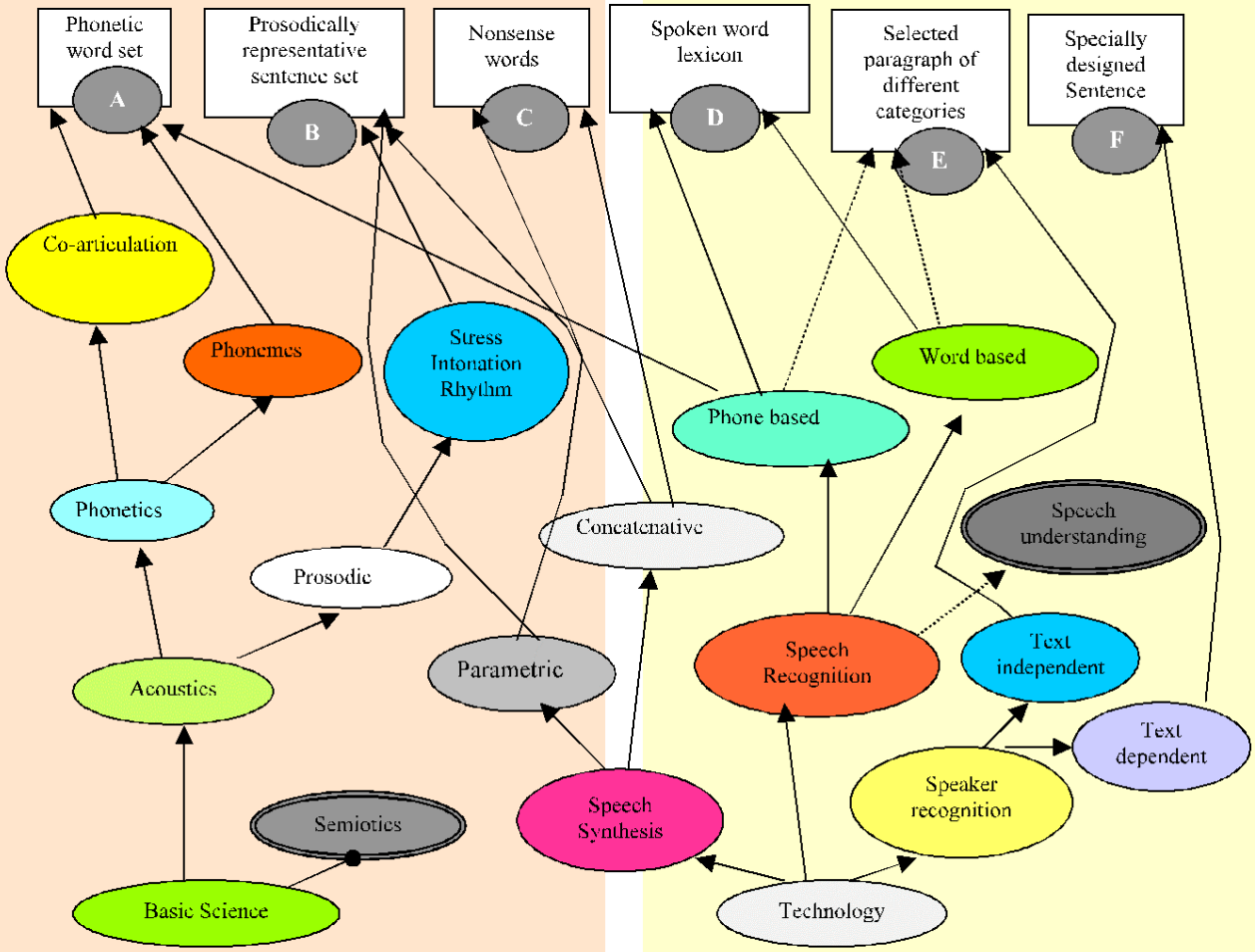


*Figure 1: Considerations for Content Selection for Speech Corpora*

The tagged notations of the corresponding word or sentences are generated as per the corpora management structure. For convenience, the text and associated speech contents are maintained under the same filename with different file extensions. While the speech recording files are maintained under a separate directory, the corpora management (backed-up with the associated user interface) had been designed in popular XML.

There is a requirement of commonly known phonetic orthographic transcriptions of entire corpora text content to permit viewing under International Phonetic Association (IPA) symbols of the orthographic representations of the same text content.

The corpora should have the two files corresponding to each of the word or sentence for a particular field, one speech sound in .wav and the other tag file with .txt file extension. The name is composed (as in Figure 2) with three component fields; namely, 1) Field ID, 2) Speaker ID, and 3) Word or sentence ID.

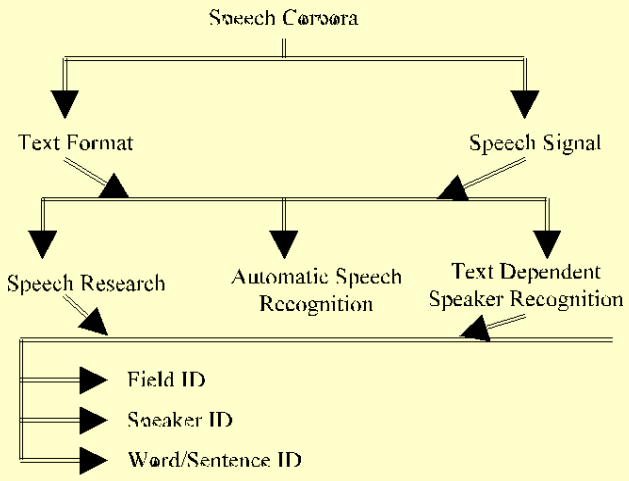


*Figure 2 Corpora management tree*

## III. POSSIBLE EXTENSIONS TO THE SPEECH CORPORA

Developed speech corpus by C-DAC, Kolkata [5] exemplifies the process of building speech corpora for Indian languages. For comprehensiveness, it may be advisable to increase the informants. The above corpus is developed for only one of the official dialect for selected languages. However, it may be necessary to build corpora covering the different popular dialects of the languages. There is also a need for special type of corpora to exemplify the speaking style of different speakers. In speech technology development (especially for Speech Recognition), the environmental noise and channel noise also add important attributes; hence, there may be a need that corpora contain sufficient samples of that nature also. The content for acoustic prosodic study in the present corpus is only for one mode of talking. The corpus may need to be extended to include other modes (like anger, happiness, joy etc.) Figure 3 describes the possible Extension of the developed basic speech corpora.

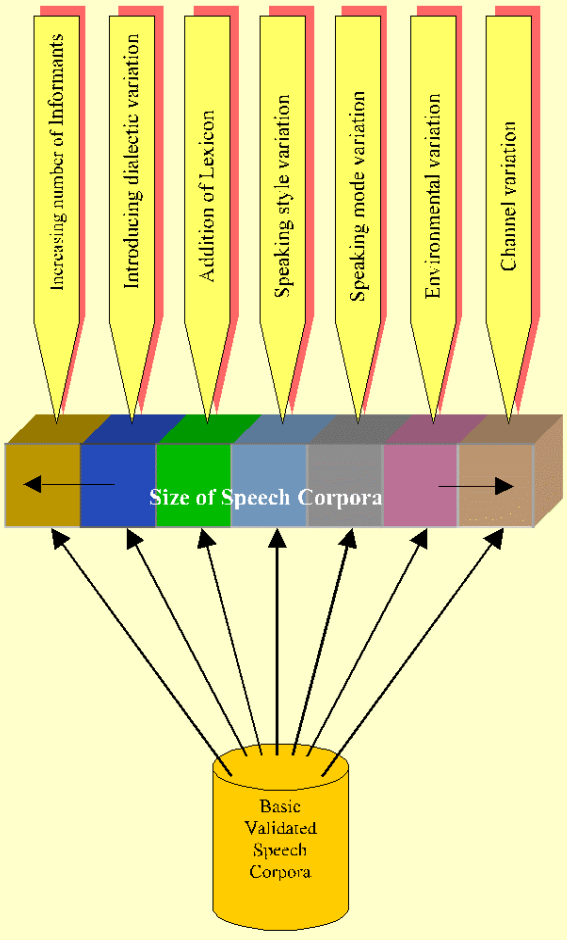


*Figure 3: Possible extensions of speech corpora*

To develop the channel variation and environmental noise data based for speech recognition technology, existing recorded content is rerecorded through different channel and environmental noise. The developed content is already annotated. So, if we use the same content, then same annotation can be useful after recording through different channel and environmental noise. Only initial time alignment has to be made. Following channel variations are incorporated in the existing speech corpora content for speech recognition and speaker recognition technology development.

- Mobile-to-Mobile: The existing speech studio condition recorded content is transmitted through a mobile and after receiving by a mobile it is recorded in computer.
- Mobile to Land Line Phone: in this case the transmitter is Mobile phone and receiver is a Land Line Phone
- Land Line Phone to Mobile: in this case the transmitter is Land Line Phone and receiver is a Mobile phone
- Land Line Phone to Land Line Phone: in this case both the transmitter and receiver are Land Line Phone.

All the modulation technology like FDMA, TDMA, and CDMA are covered.

In case of environmental noise the above corpora content is recorded from railways station, domestic airport, busy road, and market place.

## IV. CONCLUSION

The aforesaid Speech Corpora is considered essential to support the Speech Technological research and development for man-machine-man Interface in Indian Languages. This Corpus may appear to be adequate only within the limited context of its design. Primary restriction is towards the manner of speaking (i.e. Standard Text Reading Mode). The database for the study of supra-segmental characteristics may be inadequate for deriving rules for the purpose of generation of fully natural speech. It may be extremely difficult to mimic spontaneous human speech mechanically. Spontaneous interactions between speakers are not yet included in the present corpora. Furthermore, emotionally rich contents are deliberately avoided. In future, speech corpora containing normal spontaneous emotional speech may be designed for both (a) broad based spoken language processing (SLP), as well as (b) linguistic studies on speech. Towards this enlarged objectives, such speech corpora may also provide the following:

- Normal spontaneous dialogue
- Natural query-answer on various topics of importance in the selected domain.
- Inclusion of dialogues depicting various emotions (data could be collected inter alia from creations of performing arts like Cinema, Television, etc.)
- Inclusion of spontaneous speech (memorable events in the life of informants).

## REFERENCES


[1] European Lang Resources Assoc. http://www.icp.grenet.fr/ELRA/.

[2] Linguistic Data Consortium. http://www.ldc.upenn.edu/

[3] Shyamal Kr. Das Mandal, A.K. Datta, "Annotated Speech Corpora Creation for Bangla", Proceedings of SIMPLE-04, IIT Kargapur, March 2004.

[4] Shyamal Kr. Das Mandal, A.K. Datta, "Annotated Speech Corpora Creation: An Approach" Proceedings of FRSM-2004, Annamli University, January 2004.

[5] Shyamal Kr. Das Mandal, Indranil Sarkar, Arup Saha and Asoke Kumar Datta, "Creation of Speech Corpora for Speech Research and Speech Technology Development: An approach", ICSLT-O-COCOSDA-2004, New Delhi, 17-19 Nov.20.

[6] Shyamal Das Mandal, Arup Saha, A.K.Datta "Annotated Speech Corpora Development in Indian languages" Vishwa Bharat vol. 16, pp49-64, January 2005.

[7] Shyamal Kr. Das Mandal, Arup Saha and Asoke Kumar Datta, "A semi-automatic speech signal annotation system" Proceedings of FRSM-2006, January 9, 2006, pp. 157-163.


*This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.*