Problems and Prospects in Collection of Spoken Language Data

Kishore Prahallad+*, Suryakanth V Gangashetty*, B. Yegnanarayana*, D. Raj Reddy+
*Language Technologies Research Center (LTRC)
International Institute of Information Technology, Hyderabad, India.
+School of Computer Science, Carnegie Mellon University, PA, USA.
Email: skishore@cs.cmu.edu, svg@iiit.ac.in, yegna@iiit.ac.in, rr@cmu.edu

Abstract **- In this paper, we focus on the information in speech data and discuss the research issues involved in collecting, organizing, indexing, retrieving and summarization of speech data. We share our experience about the problems and prospects in collection of spoken language data. We highlight some of the procedures, standards that need to be adapted in collecting the speech data, and discuss our plan to collect the speech data for all the languages in Indian sub-continent and from the neighboring countries.**

Index Terms **— Spoken language data, Speech processing, Digital libraries.**

I. INTRODUCTION

Most of the current and future digital data is in audio and video format. The increasing storage capacity, processing power and bandwidth enable us to collect such large amount of audio and video data and store it in digital form [1]. Most of this data is unstructured and unorganized. Just as the text data could be accessed quickly and easily with the support of full-text indexing and search, speech/audio and image/video data also needs to be accessed quickly by providing the query in text/speech/audio/image/video modes. Organizing, indexing, retrieving and summarization of large corpora of speech/audio and image/video data pose challenging research issues for engineers and scientists as most of the future digital libraries consist of audio and video data only [2].

In this paper, we focus on information in speech data and discuss the need for collection of spoken language data. We share our experience about the practical and research issues and prospects in collecting the spoken language data.

This paper is organized as follows: The characteristics of spoken language data are discussed in Section II. Section III describes the classification of spoken language data. The research issues in processing the spoken language data is discussed in Section IV. Section V discusses the earlier effort in the collection of spoken language data. The various tools developed and used for the data collection tasks are described in Section VI. Further extension of data collection tasks for all the Indian official languages is described in Section VII. Final section describes and some of the conclusions.

II. CHARACTERISTICS OF SPOKEN LANGUAGE DATA

The term audio refers to a class of sounds which is not produced by human speech production mechanism. Typical examples of audio data are sounds produced by instruments, machines,

nature and its species other than the human-being. Speech is a class of sounds produced by human speech production mechanism. It includes meaningful sentences, short words and non-meaningful sounds such as whispers, humming, whistling etc.

The spoken form of a language bears more information than its written / text form. The fact that the speech is produced by human production mechanism, it carries the signature of the speaker and his/her state, message to be transmitted, language etc. Along with the message (information), speech data contains information about the following:

1. Speaker – Who is the speaker?
2. His/her background – Age, gender, literacy levels, knowledge levels, mannerisms etc.
3. Emotions – Anger, sad, happy etc.
4. Idiolect – An individual distinctive style of speaking
5. Medium of transmission – Microphone, telephone, satellite etc.
6. Environment - The information about the surroundings of the speaker such as party-environment, airport/station, office, quiet space etc, is also being captured along with the spoken data.
7. Language.
8. Dialect – The distinctive grammar and the vocabulary associated with a regional or social use of a language.
9. Culture and civilization – The richness of usage of vocabulary, grammar etc, indicates the times of the language and the society.

Some more interesting characteristics of spoken language data are associated with its study over a period of time. It would be interesting to know or able to answer the following:

1. How a language was spoken 25 years ago, 50 years ago, 100 years ago and beyond?
2. How a famous poem was recited or sung by the author?
3. How a particular language was spoken in different geographical locations of a state/country?
4. How a particular language/dialect has evolved over a period of time?
5. What were the rare languages/dialects (which were no more in existence)?. How they were spoken?

Collection, storage, indexing and retrieval of spoken language data are essential as the spoken language seems to embed multidimensional facets of a person, language, society and the civilization itself. The other useful purposes for collecting the spoken language data are as follows:

1. Information access by the illiterate, visually impaired persons through speech-mode. The collection of stories, lectures, blogs may serve as useful resources for information dissemination.
2. Promotion of research and development in the areas of speech and language.
3. Development of speech systems by making use of collection of large speech corpus.

III. CLASSIFICATION OF SPOKEN LANGUAGE DATA

Spoken language data could be classified based on the mode of speech, medium of recording, language etc.

(a) Mode*:* Some of the different forms/modes of spoken language data are short-stories, stories, novels, lessons, lectures, seminars, meetings, conversational, party, blogs, counseling, religious and political speeches, nation addresses and traditional songs.

(b) Medium*:* Microphone, multi-array microphone, broadcast medium such as TV, radio, studio quality recording as in movies, telephone, mobile-phones, cam-coders, wireless devices used in military and defense applications

are some of the mediums through which the speech data could be collected.

(c) Environment*:* For several research purposes it may be necessary to collect speech data in different environments and backgrounds such as cock-tail parties, noisy environments such as super-markets, offices and car-noise. Such collection of data could be used to estimate the characteristics of different types of noise which could be useful for speech enhancement, speech recognition and other applications.

(d) Languages and Dialects*:* Jim Baker's proposal is to collect one million hours of speech in all languages and dialects of the world. As a part of this global project our efforts are focused to collect spoken language data in languages of India and its neighboring countries.

(e) Non-native speech*:* Another class of speech is the non-native speech such as English spoken by Indians. Such collection would be useful to study the influence of English on Indian languages in terms of pronunciation, usage of vocabulary etc.

(f) Parallel-Corpus*:* Existence of parallel spoken language corpus is two or more languages assume relevancy to do research and build models such as speech-speech translation systems.

(g) Multimedia-Data*:* Here the speech data is captured along with visual information about the speaker. Such data aids in study of relationship between speech and visual media and vice-a-versa.

IV. RESEARCH ISSUES IN PROCESSING THE SPOKEN LANGUAGE DATA

(a) ***Speech without transcription****:* Often speech data such as conversational data, lectures, movies, drama, meetings, discussions etc, are often captured in audio and or video format through specialized equipment, desktops or small-devices such as mobile phones. Note that a large amount of such data is captured without the transcription (the text-form of the spoken-data) which is essential for current generation algorithms to index and retrieve the audio-data. There is essential need for sophisticated algorithms to process speech data without any transcription and perform operations such as segmentation, indexing, retrieval and summarization. Algorithms in the direction of automatic derivation sound units from the acoustic data, or which use prior knowledge such as Global phone set with adaptation techniques may be useful.

(b) Unit for Indexing*:* The indexing of a text corpus is done by using inverted file index format where words are used as units. One way is to index the speech data is also at the word-level or sub-word level provided we have the transcription or could generate the transcription. Other methods of indexing the speech data such as using automatically derived units, prosodic information, acoustic-phonetic hints, global phone set, could also be explored.

(c) Querying in text or speech mode*:* A user may key-in a set of words to retrieve the corresponding speech data or may speak out the query. Note one of the issues would to map the text to the units of indexing. This mapping could be complex if the units of indexing are derived from acoustic data.

(d) Retrieving*:* Given a query the user would like to retrieve a result or a set of results and

have ranked in some order of relevance. Given a query in speech mode, one of the classical issues would be design of the algorithms for fast matching of acoustic hints and accurate retrieval on a large corpus.

***(e) Summarization**:* The speech data which is retrieved could be longer in terms of duration and may have to be played out to the user. User often may be more interested to have a summary of a large speech file as apposed to listening to the entire speech data. A simple approach is to work with conventional text-summarization approaches by working on transcription of speech data, however, other approaches may involve development of models by extracting different features at different levels such as prosody-level and semantic-level.

***(f) Speaker-Segmentation**:* The retrieved data could consist of speech involve two or more number of speakers, and a user may be interested on a particular speaker. Multi-speaker segmentation using supervised and un-supervised techniques have to be employed to segregate the speech of a speaker.

***(g) Speech-Enhancement**:*
The retrieved data may be noisy inherited from its original recordings. The quality of such speech signal needs to be enhanced before playing back to the user.

***(h) Language Identification**:*
One of the attributes for indexing as well as for retrieval could be the language of the spoken data. Automatic identification of the language is essential using small amounts of speech data in real-time.

***(i) Cross-Lingual Information Retrieval**:*
It is often possible that the relevant result available in other languages, or the user may want to search across languages. Such approaches may be possible if the indexing being done at the phonetic level, and the user keys in what sequence of sounds he/she is looking for.

In case of India, the languages share a common phonetic base, which essentially makes it easier to query or search. Often the result retrieved from another language could as well be played without any translation, as most of the Indians are found be bilingual.

***(j) Speech-Signal Processing**:*
Most of the feature extraction algorithms in the area of speech rely on second order statistics methods such as autocorrelation, Fast-Fourier transform, linear prediction etc. In view of many of the issues mentioned above it is necessary to explore new methods of processing and extracting of features from the speech signal which capture higher order statistics and exploit non-linear relationship among the samples.

V. EARLIER EFFORTS ON SPEECH DATA COLLECTION

In this section, we will discuss some of the speech data collection efforts at IIIT Hyderabad (IIITH) and IIT Madras (IITM)

VA. Speech Data Collection at IIIT Hyderabad

The speech data collection at IIIT Hyderabad was done in two different modes: (a) High-quality studio recording, and (b) telephone and cell-phone speech.

***(a) High Quality Studio Recordings**:*

The high quality studio recordings were single-speaker databases of 1-2 hours, and were recorded by native speaker in Telugu, Hindi, Tamil, and Indian-English. These recordings are read speech spoken naturally but speakers haven't added any specific emotions while reading. The speech signal was originally

recorded at 44 KHz with 16 bits per samples, but was down sampled to 16 KHz with 16 bits per sample. These databases are used for building unit selection based speech synthesis systems [4].

(b) Telephone and Cell-Phone Corpus:

The telephone and cell-phone speech corpus was recorded by 540 speakers in Tamil, Telugu and Marathi languages over landline and cellular phones using a multi-channel computer telephony interface card. A speaker from any where in India could call the telephone connected to a computer in speech laboratory at IIIT Hyderabad and could start recording the speech. The speech signal was sampled at 8 kHz with 16 bit resolution.

Speech data was collected from the native speakers of the language who were comfortable in speaking and reading the language. Speakers from various parts of the respective states (regions) were carefully recorded in order to cover all possible dialectic variations of the language. Each speaker has recorded 52 sentences. Table 1 gives the number of speakers recorded in each language and in each of the recording modes – landline and cellphones.

Table 1: Number of speakers in the three languages

Language	Landline	Cellphone	Total
Marathi	92	84	176
Tamil	86	114	200
Telugu	108	75	183

Despite the care taken to record the speech with minimal background noise and mistakes in pronunciation, some errors had crept in while recording. The transcriptions were manually edited and ranked based on the goodness of the speech recorded. The utterances were classified as “Good”, “With channel distortion”, “With background noise” and “Useless” whichever is appropriate. The pronunciation mistakes were carefully identified and if possible the corresponding changes were made in the transcriptions so that the utterance and the transcription correspond to each other. The idea behind the classification was to make the utmost utilization of the data and to serve as a corpus for further related research work.

Using these speech databases large vocabulary speech recognition systems in Telugu, Tamil and Marathi were built [11]. The performance of these speech recognition systems is shown in Table 2.

Table 2: Performance of speech recognition (ASR) systems in %WER (Word error rate)

ASR System	WER (%)	Vocabulary
Marathi Landline	20.7	21640
Marathi Mobile	23.6	18912
Tamil Landline	19.4	13883
Tamil Mobile	17.6	16187
Telugu Landline	15.1	25626
Telugu Mobile	18.3	16419

VB. Speech Data Collection at IIT Madras

This speech data collection was done at speech and vision laboratory at IIT Madras. TV news bulletins broadcast by Doordarshan in Tamil, Telugu and Hindi languages were recorded on the same day so that the semantic content of the news bulletins is nearly the same. This helps in acquiring many examples of different nouns and pronouns in different languages, and for audio content analysis independent of the language. The audio signal was sampled at 16 kHz with 16 bit resolution, and stored as separate news bulletins.

Each bulletin represents approximately 10-15 minutes of read speech. Thus 33 bulletins spoken by 10 male and 23 female speakers for Tamil language, 20 bulletins spoken by 11 male and 9 female speakers for Telugu language and 19

bulletins spoken by 6 male and 13 female speakers for Hindi language were collected. The total duration of the speech data is about 5 hours for each of the three languages. The read speech was manually segmented into short segments of approximately 3 sec duration, containing only the news-readers speech. Speech by other speakers in the bulletin was ignored since the context and style used could be different. Speech utterances corrupted by channel or other noise were clipped out. The segments were then orthographically transcribed manually into a common Indian Language Transliteration (ITRANS) code (Roman script) for Indian languages [5]. The ITRANS code was chosen, as it uses the same symbol to represent common sounds across the Indian languages. This forms a good intermediate script to compare, contrast and represent the sound units of different languages. Allophonic and other variations were maintained in the transcription using the closest ITRANS code. The transcribed sentences were parsed into syllables based on the rules of the language. The speech segments were then segmented into syllables manually by trained human subjects. The database was organized in a TIMIT-like format [6]. Each bulletin in the speech database is organized into separate directories. Each directory containing the original wave files, the transcription of the speech file, time aligned segmentation of the speech file into words and syllables.

This speech data base and its information is utilized for the following speech processing areas:

1. The statistical information of the distribution of the syllables in terms of their frequency of occurrences and their durations in the design of limited vocabulary speech-to-text conversion system [7][8].
2. Text-to-speech synthesis [9].
3. Language identification system [10].
4. The usefulness of the knowledge of relative frequencies in the context of HMM models for speech recognition, and spoken language identification.
5. The importance of duration modeling for HMMs and Text-to-speech synthesis.

The above studies suggested that, the statistics from a large speech database in Indian languages can be used to improve the performance in speech recognition, speech synthesis and language identification.

VI. TOOLS TO AID THE SPEECH DATA COLLECTION EFFORTS

The following are some of the tools that were built and refined during the data collection efforts.

(a) Recording Tool: This is for general purpose recording on the desktop, where the sentence to be recorded would be displayed along with record options. Upon recording, it will display the waveform to ensure that the speech has been recorded and it is not clipped. The interface also provided options for saving, navigation to the next, to previous or to any sentence.

(b) Computer-Telephony Recording Tool: This tool uses dialogic (or Intel) card that facilitates computer-telephony interaction. Software has been built using this card that facilitates the user to record speech by calling this computer from a typical land-line or cell-phone. In telephonic recording some of the issues are (a) How to implement start and stop (automatic or manual), (b) How to navigate to next or previous recording, and (c) How to navigate to any sentence. This software handles these issues seamlessly by allowing the user: (a) to press * to start recording and press * to stop recording, (b) to press # (and sentence number) to navigate to any sentence for re-recording.

(c) Transcription Correction Tool: Typically a read speech contains the following error types: (a) insertion, deletion or substitutions

at phone, syllable or word level, (b) interference of background noise or other speakers' voice, (c) abrupt begin or end, (d) unclear pronunciation, and (e) very low speech-to-noise ratio. The purpose of transcription correction tool is to aid the correction in three different phases. In phase 1, this tool allows the operator to give a binary decision of speech being same as the expected transcription or not. Phase 2 deals with the errors of type 1 and 2 which requires correction of the transcription. Phase 3 deals with the error type 3 which involves cutting of the speech signal.

(d) Audio and Video Transcription Tool: The purpose of this tool is to key-in the transcription by listening to the speech present in movies, recorded lectures etc. This tool allows short keys to stop, back-off to the nearest breath pause etc, through which an operator can key-in the text much faster than by using the media-player.

(e) Tool for Correction of Segment Labels: Given the transcription and the speech data, labeling at segment level can be generated by using force-alignment of hidden Markov models (HMM). However, the automatic labels needs to be checked and corrected for accurate labeling. Emulabel is such a tool which assists in correction of speech labels in easier and faster way. This tool allows the user to simply drag the label-marker to the required position thus reducing the correction time.

VII. EXTENSION OF DATA COLLECTION FOR ALL OFFICIAL LANGUAGES OF INDIA

It could be observed from the efforts at IIITH, IITM and also other engineering and linguistic/phonetic departments that these are isolated efforts of speech data collection. Moreover, the collected speech data is typically smaller is size: – 90 hours (540 speakers in 3 languages) by IIITH and 15 hours (72 speakers in 3 languages) by IITM. While there are some common processes and tools used for speech data collection but these efforts need a unified framework where a set of procedures, processes and tools are standardized. Also the speech data collected at each center or university is typically limited to a restricted use as opposed to available to public for general research purposes.

Our future plan is the following:

1. Standardize the speech data collection process, procedure and tools.
2. Collection of speech data about 100000 hours (.01 million hours) for each official languages of India.
3. There are 23 official languages in India, and the collected speech data would cover roughly 20% of the Jim Baker's global project-proposal of collecting 1 million hours of speech in all languages of the world.
4. This speech data collection would include collection of all classes as discussed in Section III.
5. Organize, annotate and index this speech data.
6. Make available this data for general public use and for research purposes.

We believe the impact of such large data collection could be as follows:

1. Availability of information in spoken language form for illiterate and others.
2. Promotes research in speech technology for Indian languages.
3. Enable to develop speech technology products useful for common man.
4. Examples of speech technology products include speech-speech translation systems for information exchange, screen readers for illiterate and physically challenged, naturally speaking dialog systems for information access over voice mode.

VIII. SUMMARY AND CONCLUSIONS

In this paper, we have discussed the characteristics of spoken language data and the

need for collecting spoken language data. We have described the speech data collection efforts at IIITH and IITM. The procedures, tools used in this data collection and the outcome of these efforts are also discussed. We have also shared our plan and the need for collection of larger corpus covering all official language of India under a unified framework.

IX. REFERENCES

[1] Gordon Bell, IEEE Spectrum, November 2005.

[2] Swcrisis, IEEE Spectrum, September 2005.

[3] Zhu, Zhu, Lee, Simon, *Cognitive theory to guide curriculum design for learning from examples and by doing*, Journal of of Computers in Mathematics and Science Teaching, 2003, 22(4), 285-322.

[4] S. P. Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal, *Experiments with unit selection speech databases for Indian languages*, National Seminar on Language Technology Tools: Implementations of Telugu, Hyderabad, India, 2003.

[5] ITRANS- Indian Language Transliteration Package Version 5.2 Source http://www.aczone.com/itran/itrans.

[6] William M. Fischer, George R. Doddington and Kathleen M. Goudie-Marshal, *The DARPA speech recognition research database: specifications and status*, in Proc. DARPA workshop on Speech Recognition, pp.~93—99, February 1986.

[7] A. Nayeemulla Khan, S. V. Gangashetty, and S. Rajendran, *Speech database for Indian languages - A preliminary study*, in Proc. Int. Conf. Natural Language Processing (Mumbai, India), pp. 296--301, Dec. 2002.

[8] A. Nayeemulla Khan, S. V. Gangashetty, and B. Yegnanarayana, *Syllabic properties of three Indian languages:~Implications for speech recognition and language identification*, in Proc. Int. Conf. Natural Language Processing (Mysore, India), pp.~125--134, Dec. 2003.

[9] K. Sreenivasa Rao, S. V. Gangashetty, and B. Yegnanarayana, *Duration analysis for Telugu language*, in Proc. Int. Conf. Natural Language Processing (Mysore, India), pp.~152--158, Dec. 2003.

[10] Leena Mary, K. Sreenivasa~Rao, S. V. Gangashetty, and B. Yegnanarayana, *Neural network models for capturing duration and intonation knowledge for language and speaker identification*, in Eighth Int. Conf. Cognitive and Neural Systems (Boston, USA), p. 9, May. 2004.

[11] Anumanchipalli Gopalakrishna, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Singh, R.N.V Sitaram and S.P. Kishore, *Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems*, in Proceedings of International Conference on Speech and Computer (SPECOM), Patras, Greece, Oct 2005.