

Multilingual Crowdsourcing Methodology for Developing Resources for Under-resourced Indian Languages

Karunesh Kr. Arora¹, Sunita Arora¹, Mukund Kumar Roy¹, Shyam S. Agrawal²
CDAC¹, KIIT²

C-56/1, Sector 62, Noida, India¹, Sohna Road, Gurugram, India²
{karunesharora, sunitaarora, mukundkumarroy}@cdac.in, ss_agrawal@hotmail.com

Abstract

Huge Data collection challenge gets intensified for Under Resources Languages especially for large variety of Indian languages. We propose building of a common framework for collecting, monitoring and evaluating speech resources irrespective of the language. Common phone set for transcribing and annotating the data, easy portability, configuring different languages, indigenous algorithm for extracting phonetically rich sentences, on-line and off-line recording facility, tacking code-mixed data, quality checking & control are unique features that enable the framework collecting data from remote and rural parts of the country. It also highlights issues and hurdles faced in collecting sample data and addressing them.

Keywords: Crowdsourcing, Under Resource language, Phonetically Rich sentences

1. Introduction

Current language and speech technologies are data driven. This is due to the fact that model built over data needs to collect evidences from different instances. Large amounts of annotated speech data are needed to model the effects of different sources of variability. An axiom of speech research is - there are no data like more data.

India is one of the largest and fastest growing markets for digital consumers, having 560 million internet subscribers in 2018 (TRAI, 2018), second only to China. According to McKinsey report 2019, India is one of the largest and fastest growing markets for digital consumers, and India's lower-income states are bridging the digital divide, and the country has the potential to be a truly connected nation by 2025. The experiment described in this paper presents exploiting the use of smartphones for collecting huge and varied speech data. The speech data collection has crossed the boundaries of studio based sequential recordings to crowd-source based parallel recordings. This paper details such a framework which can be used for multiple languages, utilizing common phone-set for Indian languages, indigenous algorithm for extracting phonetically rich sentences, on-line and off-line recording facility, tacking code-mixed data, quality checking & control while recording, and enabling collection of voice data from rural and remote areas at the user's choice of time in multiple sessions.

2. Related Works

Switchboard (Godfrey et al., 1992)[1], Fisher (Cieri et al., 2004)[2] and Broadcast News (Garofolo et al.,1997) corpora[3] are some of the prominent high-quality corpus. These all have taken a long time and huge effort and are considered landmark in speech technology field.

In Indian languages Speech Database for Hindi, Indian English and Bengali languages have been recorded by 1500 speakers in each language covering different environment conditions, Age Groups and gender distribution. The speech data is collected through IVR

mechanism over mobile recorded speech data, sampled at 8 KHz. The crowd-sourcing mechanism has been is use for sometime Crowdee, CowdFlower etc[4,5]. The effort and framework presented here advocates and uses an indigenous corpus design methodology, yet provides the advantage of speedy and near to real life scenario data collection which is desired for building robust ASR system. The audio recordings are not limited to some specific microphones or recording devices. This also helps in collection of speech corpus covering a variety of recording devices and thus generalizes the speech corpus.

3. Architecture

The client application works on the user's Smartphone. The overall communication happens over the HTTP protocol. The Android based App 'मेरी भाषा मेरी वाणी' [6] (My Language My Voice) facilitates user to read out the sentence or phrase displayed over the screen. The session is maintained once a user is connected to server. On completion of recording, the user has the option to verify it through playing it before final submission. On pressing submission button, the preliminary level validation is carried out on the client side and on passing the preliminary check the recording is submitted to the server and next prompt is received. The App is Android based and works on majority of smartphones being used in India.

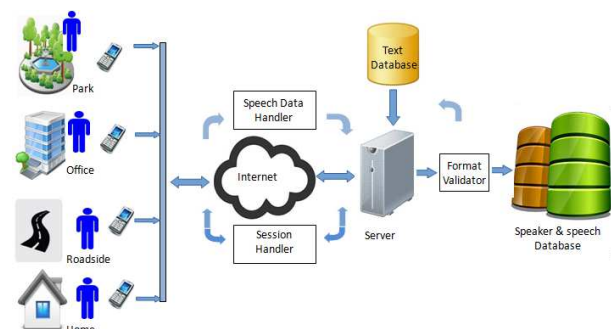


Fig. 1. Client-Server architecture for speech data collection

4. Text Corpus Development

Cleaning and filtering of corpus: The html boilerplate are removed from documents. The text is filtered on the basis of character encoding (UTF-8)/ASCII to adhere to the consistency of English alphabets and numerals. Additional portal specific contents for example URLs, emails, addresses etc. are dropped. Identification of improper syntax e.g. existence of invalid bigrams/character combinations has been done. Sentences with foreign word presence are filtered so as to have a good quality monolingual corpus. Inadequate sized sentences and words are identified and removed. Invalid Unicode patterns in sentences and Out of range character in sentences are not considered. Duplicate sentences along with duplicate punctuations are also removed.

An algorithm was developed to extract Phonetically/Grapheme Rich Sets from any corpus. The thresholds for the same was made configurable in order to limit sentences. The grapheme ratios were maintained to be a factor of original representative corpus. Maintaining such ratios provided a more natural way of increasing phonetic richness of corpus.

Buckets creation: In order to include the extracted phonetically rich sets a different algorithm was worked on

- The sentences were categorized into buckets based on sentence lengths
- The buckets were sorted according to their grapheme richness.
- Also, a separate data structure was maintained in order to a track each grapheme present in a sentence along with the counts.
- Once the data structure is ready, algorithm adaptively decides the rare graphemes, and starts to pick one sentence each bucket.
- Works on to extract sentences till a specified threshold of all graphemes is reached.
- After that, works only to enrich the set of graphemes that are rare.
- Thus, ensuring proper distribution of sentences across the prompt sheets, taking into consideration readability and grapheme richness.

5. Components of a general prompt sheet

The database contains sentences, guided prompts and unguided prompts/queries. The sentences are designed having minimum phonemic distribution and falling within certain length. Phonetically rich sentences of length less than 16 words [5] are collected using greedy algorithm and proof reading is done to ensure correctness of data. Some sentences were framed manually and are introduced to optimize the set. The Table 1 lists the items covered in the speech database.

Speech Database contents
Phonetically rich sentences collected from news papers, books, BTEC sentences and web
Proper Names - Indian males & females, Cities/States/Countries
Visiting Places, Monuments, Parks, famous buildings, Airport names, Airline names, Railway Stations names, Train names etc.
Unguided Queries with expected responses – <ul style="list-style-type: none"> - Isolated Digits - Connected Digits - Date & Time Vocabulary - Vocabulary - Currency and Money - Measurements - Yes/No utterances
Guided Prompts containing – <ul style="list-style-type: none"> - Isolated Digits - Connected Digits - Date & Time Vocabulary - Currency and Money - Measurements Yes/No utterances
Silence to capture background noise

Table. 1. Speech Database contents

The digits and numbers vocabulary covers telephone/mobile numbers, PIN codes, credit card numbers and natural numbers, date and time expressions contain - months, days, holidays, time, Proper names contain (person names and geographical entities like cities, states and countries). Guided prompts list out the entities in word format to get recordings in the planned way and unguided prompts give freedom to the speaker to speak in natural manner. Unguided prompts appear before the guided prompts to avoid biasing. For example, in normal scenario, the mobile numbers or monetary values may be spoken by a person in more than one way. In many real cases, it can be easily observed that people even do the code switching of the language also. In un-guided scenario, it has been tried out to have the queries, so that person answers them in his/her preferred/casual manner. While in guided scenario the speaker is provided the way in which he/she is supposed to speak.

6. Quality Control during recording

The client application also performs some of the quality checks while recording itself. These include presence of short silence zone before the start of a prompt and after the end of a spoken prompt which is to clipped utterance of prompts. The second quality check is SNR ratio, if it is not above a threshold the user is asked to either move to some lesser noisy place or speak a bit louder.

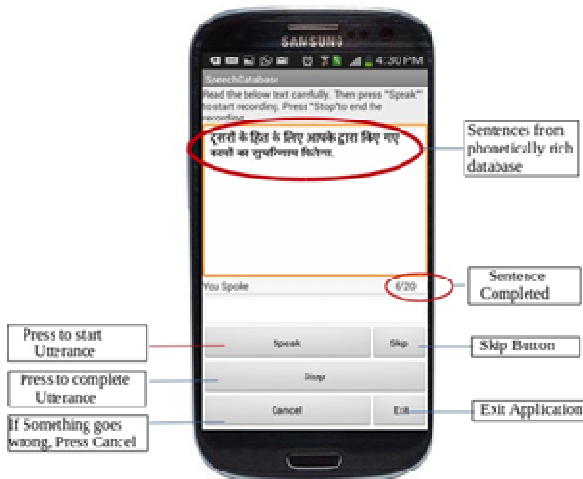
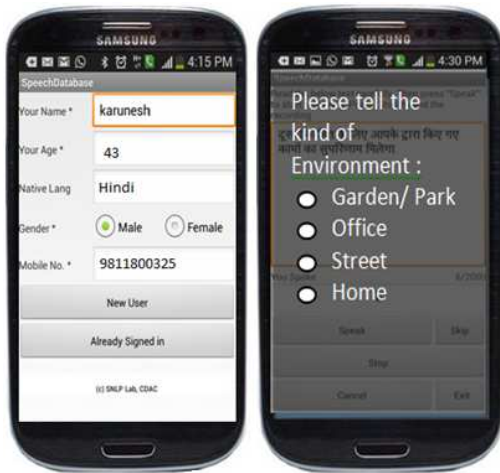
7. Interface for recording

The speaker is provided with a user friendly interface on the mobile. The Registration phase captures the Meta information about the speaker, such as his demographic details, environmental details and equipment information.

7.1 Client Application

Client application is a mobile app facilitates user with an interface for recording voice as well as on screen display of sentences to utter. Before initiation of the recording session, the user has to register. As part of the registration, user provides Meta information like Name, Age, Gender, Mother Tongue, Native Place, Mobile Number and environment in which they are currently accessing the App (Fig. 2).

After Registration phase, the recording session starts. The sentences/prompts keep on appearing on the mobile screen, one after another, and user is also provided facility to skip the sentences/prompts he/she feels uncomfortable and can repeat recording if not satisfied. The speaker can record in single or multiple sessions as per his/her convenience (Fig. 3). The sentence/prompt wise speech data is transferred to the server, as soon as he/she goes for the next.



App has two modes of operations:

- Online mode:** In this mode, speaker's recording is sent to server upon completion of each utterance, and next sentence appears only when transaction is successful.
- Offline mode:** In this mode, where internet connection is not available, user is presented with pre-stored set of sentences in the App. He can record his utterances one by one and complete his session. Later on, whenever there is internet facility is available for him, the recordings can be uploaded in batch.

8. Recording Workflow

Here, we list out the steps being followed in Recording process.

8.1 Selection of speaker

Though the App is accessible to all, yet in initial phase, we have provided access to various institutions where a number of speakers can contribute their voice. For this institute level login facility has been extended for managed crowd-sourcing.

8.2 Registration and Meta-data information

A basic electronic record of the speaker's personal information is entered into the mobile device, including Age, Mother Tongue, Place and level of education, mobile number and data collection consent etc.

8.3 Training

The speaker is briefed and trained to utter different prompts after indication beeps etc. Sometimes, sample recordings are done. Videos / Audios of pre-recorded sessions are played. First two prompts are provided for the purpose of training and getting feel of the whole process.

8.4 Recording

Actual prompts recording takes place after successful training and confidence level of speaker.

8.5 Reward

Upon completion of the recording session, the primary validation of recorded voices is also carried out to ensure the proper recordings. The respondent is rewarded as per prior agreement.

9. Observations & Challenges

The most of the recordings were found by the speakers in the age-group of 18-30 years. Some of the issues observed are listed below:

- Speaker sometimes moved mobiles away from mouth (to read the prompt from mobile screen) while recording, so intra-utterance variation in amplitude was observed.
- However, not fumbled, but broken pronunciation of difficult words was observed like फास्फोलिपिड्स, कार्बोहाइड्रेट-उपापचय, डाइसल्फाइडेज.
- Long silences in-between and at the end of sentences captured unnecessary noise.
- Some speakers' recording were not in natural as expected, though this ratio is quite low.

10. Future Work

The paper presents here the use of crowd-sourcing through the most common device mobile for speech data collection in collaborative and cost effective manner. The speech data collection span got reduced. It may also help in maintaining naturalness in the speech, as people felt more comfortable speaking to the mobile. The system comes with easy to use interface and prompts the speaker with sentence/text to be read/spoken by him/her. The bar above the display window provides the instructions to guide the speaker. This helps in simultaneous recordings with minimum manual handholding. The speaker is able to complete his recording in split sessions. In speech database collection through this way, we do not have any control on microphone type; distance of phone, ambient noise etc., yet this comes with the advantage of gathering the speech data in close to natural way of speaking. The data collected in this way is representative of the actual test data which ASR would be subjected to in the web or mobile based application. As a future work this data would be used to improve the acoustic models of the existing ASR system [7]. The framework used in this experiment is configurable and independent of the language and will also be used for other languages.

11. Conclusion

It is time to save cultural heritage of language and associated culture in India. There are many languages which are almost zero resourced. New tools and technologies combined with innovative approach to attract Crowdsourcing can really help to revive many endangered languages. Through "मेरी भाषा मेरी वाणी", we have planned to create Speech resorces for Punjabi, Bengali, Assamese and Gujarati in the first phase. This languages are less-resourced to start with. Later on we will cover very less or zero resourced languages.

12. Acknowledgements

We thank the speakers who have contributed their speech data. Thanks are due to Mr. Dipankar Ganguly for developing algorithm for extracting phonetically rich sentences. We thank our Executive Director and management for providing conducive environment for performing this task.

13. Bibliographical References

1. Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In Proceedings of Acoustics, Speech, and Signal Processing (ICASSP-92). IEEE.
2. Cieri, C., Miller, D., and Walker, K. (2004). The Fisher Corpus: a resource for the next generations of speech-totext. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC). European Language Resources Association.
3. Garofolo, J., Fiscus, J., and Fisher, W. (1997). Design and preparation of the 1996 Hub-4 Broadcast News benchmark test corpora. In Proceedings of the DARPA Speech Recognition Workshop
4. <https://app.crowdee.de/>
5. <https://data.world/crowdfunder>
6. <https://play.google.com/store/apps/details?id=in.cdac.crowdapp>
7. Sakriani Sakti, Michael Paul, Andrew Finch, Shinsuke Sakai, Thang Tat Vu, Noriyuki Kimura, Chiori Hori, Eiichiro Sumita, Satoshi Nakamura, Jun Park, Chai Wutiw WATCHAI, Bo Xu, Hammam Riza, Karunesh Arora, Chi Mai Luong, Haizhou Li, "A-STAR: Toward Translating Asian Spoken Languages", Computer Speech and Language, Special Issue on Speech-to-Speech Translation, Volume 27, Issue 2, pages 509-527, 2013. Indian telecom services performance indicator report, June-September 2018, Telecom Regulatory Authority of India.
8. Marge, M., S. Banerjee, A. I. Rudinicky, "Using the Amazon Mechanical Turk for Transcription of Spoken Language", *In Proc. of ICASSP*, 2010.
9. Yang, Z., B. Li, Y. Zhu, I. King, G. Levow, H.M. Meng, "Collection of user judgments on spoken dialog system with crowdsourcing", *In Proc. of SLT*, 2010.
10. Lane, I., M. Eck, K. Rottmann, A. Waibel, "Tools for Collecting Speech Corpora via Mechanical-Turk", *In Proc. of Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.
11. Arora Karunesh, Arora Sunita, Agrawal S. S., Paulsson Niklas, Choukri Khalid. "Experiences in Development of Hindi Speech Corpora based on ELDA standards". *In Proc. of the Oriental COCOSDA*, 2006.
12. R. Molapo, E. Barnard, and F. de Wet, "Speech data collection in an under-resourced language within a multilingual context," in 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 2014.
13. L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, pp. 85-100, Jan. 2014