Indic TIMIT and Indic English lexicon: A speech database of Indian speakers using TIMIT stimuli and a lexicon from their mispronunciations

Chiranjeevi Yarra, Ritu Aggarwal, Avni Rajpal, Prasanta Kumar Ghosh

SPIRE LAB
Electrical Engineering,
Indian Institute of Science (IISc), Bangalore, India
Overview

1. Key components
2. Introduction
3. Recording of Indic TIMIT
4. Recording setup
5. Manual annotation
6. Indic English Lexicon
7. Preliminary experiments
8. Conclusion and Future work
Key components

- **Indian English speech** corpus and Indian English lexicon.
- Approximately **240 hours** of speech recording from 80 subjects.
- Each subject spoken a set of 2342 rich stimuli.
- A set of 2342 utterances were annotated manually and obtained phoneme transcriptions to reflect learners’ pronunciation.
- Indic English lexicon containing **5,42,917** entries out of which **1,80,166** entries reflect Indian learners’ mispronunciations.
Why Indian English speech corpus?

- L2 learners’ spoken English influenced by their native language, which introduces mispronunciations and strong non-native accent.
- However, there is a lot of demand to build ASR system for non-native spoken English.
- ASR built with native English speech data are not suitable for the test conditions having non-native spoken English.
- Particularly, in CALL, mispronunciation detection and diagnosis is an important component.
- Moreover in India, English language learning has lot of demand, since it is a major language of communication in administration, law and education.
Why Indic TIMIT is unique?

Non-native Indian Context

- English speech corpus by Chinese speakers: ESCCL, SHEFCE, SELL, and SWECCL \(^a\).
- German and Italian speakers: ISLE \(^b\).
- Japanese speakers: NICT JLE \(^c\).

---


\(^b\) Menzel et al., “The ISLE corpus of non-native spoken English”, 2000

\(^c\) Izumi, Uchimoto, and Isahara, “The NICT JLE Corpus: Exploiting the language learners speech database for research and education”, 2004

Indian Context

- Most of Indian corpora were collected primarily for ASR in Indian languages.
- Few Indian English corpora exist and those do not meet the requirements of ASR.
Why Indic TIMIT is unique?

<table>
<thead>
<tr>
<th>Corpora</th>
<th># speakers</th>
<th>Duration (HH:MM:SS)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA-IICT(^1)</td>
<td>137</td>
<td>00:10:00</td>
<td>20% data have sentence transcriptions</td>
</tr>
<tr>
<td>KIIT(^2)</td>
<td>100</td>
<td>–</td>
<td>Only one speaker data is processed</td>
</tr>
<tr>
<td>IITKGP-MLILSC(^3)</td>
<td>25</td>
<td>01:22:00</td>
<td></td>
</tr>
<tr>
<td>L2-ARCTIC(^4)</td>
<td>2</td>
<td>–</td>
<td>300 sentences have phoneme transcriptions</td>
</tr>
</tbody>
</table>

- A few are limited in the number of utterances with manual annotations.
- Data collected from Indian Government organizations such LDC-IL and TDIL are limited in dialects and data size.

---

\(^1\) Patil, Sitaram, and Sharma, “DA-IICT cross-lingual and multilingual corpora for speaker recognition”, 2009

\(^2\) Agrawal et al., “Development of Text and Speech database for Hindi and Indian English specific to Mobile Communication environment.”, 2012

\(^3\) Maity et al., “IITKGP-MLILSC speech database for language identification”, 2012

\(^4\) Zhao et al., “L2-ARCTIC: A Non-Native English Speech Corpus”, 2018
Challenges in Indian context

- India is known for its language diversity, it has more than 1652 dialects/languages, out of which 22 are scheduled languages.
- It is **impractical** to record voice from the subjects belonging to all 1652 dialects/languages separately.

Strategy

- We consider languages which are scheduled languages and spoken by majority of the population.
Subject selection

- For selecting subjects, the following are considered:

- **Demographically** close languages share similar properties:
  - 1) North East, 2) East, 3) North, 4) Central, 5) West and 6) South.

- Indian languages are influenced by following **language families**:
  - 1) Indo-Aryan, 2) Dravidian, 3) Austro-Asiatic and 4) Tibeto-Burman.
## Subject selection

<table>
<thead>
<tr>
<th>Region</th>
<th>Native language</th>
<th>Population percentage</th>
<th>Originated and/or influenced language family</th>
<th>Number of subjects (M/F) recorded</th>
<th>Grouping</th>
</tr>
</thead>
</table>
| North East | Assamese | 1.28 | Indo-Aryan  
Austro-Asiatic  
Tibeto-Burman | 2 (0/) | Group-1 |
| Nepali  | 0.28 | Indo-Aryan  
Tibeto-Burman | 1 (0/1) | |
| Manipuri | 0.14 | Tibeto-Burman | 1 (1/0) | |
| Bengali | 8.10 | Indo-Aryan  
Austro-Asiatic | 8 (4/4) | |
| East  | Maithili | 1.18 | Indo-Aryan | 1 (1/0) | |
| Oriya  | 3.21 | Indo-Aryan | 3 (2/1) | |
| North  | Punjabi | 2.83 | Indo-Aryan | 2 (0/2) | Group-2 |
| Central | Hindi | 41.03 | Indo-Aryan | 14 (8/6) | |
| West  | Gujarati | 4.48 | Indo-Aryan | 4 (3/1) | Group-3 |
| Konkani | 0.24 | Indo-Aryan | 2 (0/2) | |
| Marathi | 6.99 | Indo-Aryan | 10 (5/5) | |
| South  | Kannada | 3.69 | Dravidian  
Indo-Aryan | 8 (3/5) | Group-4 |
| Telugu | 7.19 | Dravidian  
Indo-Aryan | 8 (5/3) | |
| Malayalam | 3.21 | Dravidian | 8 (3/5) | |
| Tamil  | 5.91 | Dravidian | 8 (5/3) | |

- Consider L1 spoken by ~90% population
- Group balance
- Gender balance in each group
- 16 subjects in each group
- In total 80 subjects
Subject selection

- The age of the subjects varies from 18 to 60 years with an average age of 25.42 years with standard deviation of 6.05 years.
- Subjects are UG, PG students and working staff from Indian Institute of Science, Bangalore, India.
- The subjects have variabilities in their pronunciation ability of reading English.
During, an operator carefully listen to the subject’s speech to spot any error (insertion, deletion of substitution of words).

All 2342 unique sentences from TIMIT corpus considered for the recording.

The recording is done in 16 sessions \(15 \times 150 + 92\).
Two linguistics are annotated a set of 2342 stimuli.

512, 512, 295, 511 and 512 stimuli from the five groups.
Instructions

- Use ‘∼’ to indicate the word boundaries in the phoneme transcriptions.
- In case of co-articulation between the words, merge those co-articulated words with ‘-’ symbol in the modified text box (Don’t split the words).
- Example: text transcription: “I didn’t hurt you”; the uttered phonemes in the recording “i d i d n t h ʒ tʃ u”.
- Correct: in the “Modified text” and “Phoneme spoken” boxes, the entries should be “I didn’t hurt-you”, “i∼d i d n t∼h ʒ tʃ u” but not “I didn’t hur t-you”, “i∼d i d n t∼h ʒ∼tʃ u”.

SPIRE LAB, IISc, Bangalore 13
String alignment is performed between the phoneme transcriptions in the TIMIT corpora and the manual annotated.

The percentage of correct and erroneous phonemes are comparable.

Thus, pronunciation of Indian speakers differs by a large extent from that of the native English speakers.
New word pronunciations are obtained by incorporating the variations based on mispronunciation rules to the existing word pronunciations.

### Phoneme specific context rules

<table>
<thead>
<tr>
<th>Previous phoneme</th>
<th>Target phoneme</th>
<th>Next phoneme</th>
<th>Indian specific variations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>Plosive</td>
<td>Vowel</td>
<td>Plosive is voiced</td>
</tr>
<tr>
<td>Nasal</td>
<td>Plosive</td>
<td>Any</td>
<td>Plosive is voiced</td>
</tr>
<tr>
<td>Any</td>
<td>Dipthong (except /ai/, /au/)</td>
<td>Any</td>
<td>Substituted with long vowels</td>
</tr>
<tr>
<td>Any</td>
<td>/θ/</td>
<td>Any</td>
<td>Substituted with /θʰ/ or /t/</td>
</tr>
<tr>
<td>Any</td>
<td>/ð/</td>
<td>Any</td>
<td>Substituted with /θʰ/ or /d/</td>
</tr>
<tr>
<td>None</td>
<td>Front vowel</td>
<td>Any</td>
<td>Phoneme /j/ is inserted before the vowel</td>
</tr>
<tr>
<td>None</td>
<td>Back vowel</td>
<td>Any</td>
<td>Phoneme /w/ is inserted before the vowel</td>
</tr>
<tr>
<td>None</td>
<td>/w/</td>
<td>Any</td>
<td>Phoneme /w/ is deleted</td>
</tr>
<tr>
<td>Any</td>
<td>/tʃ, dʒ, s, z,ʃ, ʒ/</td>
<td>Any</td>
<td>Substituted with /ʃs/ or /ʃz/ or /ʒz/</td>
</tr>
<tr>
<td>Any</td>
<td>/f/</td>
<td>Any</td>
<td>Substituted with /fʰ/</td>
</tr>
<tr>
<td>Any</td>
<td>/v,w/</td>
<td>Any</td>
<td>Substituted with /bʰ/</td>
</tr>
</tbody>
</table>

### Letter specific context rules

<table>
<thead>
<tr>
<th>Previous letter</th>
<th>Target Letter</th>
<th>Next letter</th>
<th>Indian specific variations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any</td>
<td>r</td>
<td>Any consonant</td>
<td>Phoneme /r/ is produced</td>
</tr>
<tr>
<td>Any</td>
<td>s</td>
<td>t</td>
<td>Phoneme /ʃ/ or /s/ is produced</td>
</tr>
<tr>
<td>Any</td>
<td>n</td>
<td>g</td>
<td>Both /ŋ/ and /g/ are produced</td>
</tr>
<tr>
<td>Any</td>
<td>r</td>
<td>None</td>
<td>Phoneme /r/ is produced</td>
</tr>
</tbody>
</table>

### Both letter and phoneme dependent context rules

<table>
<thead>
<tr>
<th>Previous letter</th>
<th>Target letter</th>
<th>Next phoneme</th>
<th>Indian specific variations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any</td>
<td>Double consonants</td>
<td>Short vowel</td>
<td>Geminate articulation</td>
</tr>
</tbody>
</table>
Lexicon construction

- Native English lexicon: augmenting all the word pronunciations from CMU, Beep and the lexicon available in the TIMIT corpus.

- A total of 3,62,751 are found in this lexicon.

- M2M aligner for letter-to-phoneme alignment\(^5\).

- Results a total of 5,42,917 entries in Indic English lexicon.

---

Preliminary experiments

ASR performance with Indic TIMIT data

Experimental setup
- Kaldi ASR tool-kit.
- Lexicon: Native English lexicon.
- Train set: 1636 stimuli from randomly chosen 63 speakers maintaining region and gender balance.
- Test set: Remaining 706 stimuli from the remaining speakers.
- LM: The sentences in the training set.
- For the comparison: trained with TIMIT train set.

Result
- WER is 15.02 with Indic TIMIT and is 93.41 with TIMIT data.
Forced-alignment performance with Indic English lexicon

Experimental setup

- Lexicons: Native English lexicon, Indic English lexicon.
- Train set: 1636 stimuli from randomly chosen 63 speakers maintaining region and gender balance.
- Test set: Remaining 706 stimuli from the remaining speakers.
- Objective measure: Phoneme error rate (PER) between estimated phoneme transcriptions and manual annotated transcriptions.

Result

- PER is 28.79 with Indic English lexicon and is 32.49 with native English lexicon.
- Erroneous pronunciations from Indian learners in Indic English lexicon help in achieving lower PER.
Conclusion and Future work

- This work describes Indic TIMIT corpus, a phonetically rich Indian spoken English corpus, to cater to the demand for large corpora under non-native speech conditions.

- This also reports the construction of Indic English lexicon, which is obtained based on the pronunciation errors made by the Indian speakers while speaking English.

- The corpus contains \( \sim 240 \) hours of speech recordings from 80 subjects and manually annotated phoneme transcriptions for a sub-set of 2342 recordings.

- Experiments are conducted to examine the effectiveness of Indic TIMIT and Indic English lexicon in comparison with the data from TIMIT and a native English lexicon.

- Further works are proposed to annotate five sets, where each set contains all 2342 stimuli from each region considering uniform number of stimuli per speaker and multiple annotators.
Acknowledgment

- To the Department of Science and Technology, Government of India for funding the project.
- The support extended by the two linguistics in the process of manual annotation.
- To all the subjects participated in the recording.
THANK YOU