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

Chiranjeevi Yarra
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
chiranjeeviy@iisc.ac.in

Ritu Aggarwal
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
rituaggarwal42@gmail.com

Avni Rajpal
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
sarkaravni@gmail.com

Prasanta Kumar Ghosh
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
prasantg@iisc.ac.in

***Abstract*—With the advancements in the speech technology, demand for larger speech corpora is increasing particularly those from non-native English speakers. In order to cater to this demand under Indian context, we acquire a database named Indic TIMIT, a phonetically rich Indian English speech corpus. It contains ∼240 hours of speech recordings from 80 subjects, in which, each subject has spoken a set of 2342 stimuli available in the TIMIT corpus. Further, the corpus also contains phoneme transcriptions for a sub-set of recordings, which are manually annotated by two linguists reflecting speaker's pronunciation. Considering these, Indic TIMIT is unique with respect to the existing corpora that are available in Indian context. Along with Indic TIMIT, a lexicon named Indic English lexicon is provided, which is constructed by incorporating pronunciation variations specific to Indians obtained from their errors to the existing word pronunciations in a native English lexicon. In this paper, the effectiveness of Indic TIMIT and Indic English lexicon is shown respectively in comparison with the data from TIMIT and a lexicon augmented with all the word pronunciations from CMU, Beep and the lexicon available in the TIMIT corpus. Indic TIMIT and Indic English lexicon could be useful for a number of potential applications in Indian context including automatic speech recognition, mispronunciation detection & diagnosis, native language identification, accent adaptation, accent conversion, voice conversion, speech synthesis, grapheme-to-phoneme conversion, automatic phoneme unit discovery and pronunciation error analysis.**

***Index Terms*—Indian spoken English data, Indic TIMIT, Indic English lexicon, mispronunciation based lexicon.**

I. Introduction

In general, most of the world population either speak or learn English. 20% of the population have English as their native language and for the remaining it is a second language (L2) [1]. An L2 learner's spoken English is influenced by their native language. This could introduce either mispronunciations or strong non-native accent in their spoken English [2], [3]. Typically, there is a lot of demand to build automatic speech recognition (ASR) system for non-native spoken English due to its demand in other applications include automatic voice response systems, computer assisted language learning, mispronunciation detection & diagnosis, virtual assistant and automatic speech translation. However, the performance of an ASR system reduces significantly when there is mismatch between the speakers' accent in train and test conditions. Thus, ASR system built with native English speech data could not be suitable for the test conditions involving non-native spoken English [4]. Further, while in ASR modeling, a lexicon containing incorrect pronunciations made by the L2 learners could be useful for handling the acoustic variabilities due to mispronunciations in English data spoken by non-native speakers.

In the literature, there exist English speech corpora from non-native English speakers. Majority of these were collected from Chinese speakers including ESCCL [5], SHEFCE [6], SELL [7] and SWECCL [8] corpora. ISLE corpus contains speech from German and Italian speakers [9]. In NICT JLE corpus Japanese speakers were considered [10]. In Indian nativity, most of the existing corpora were collected primarily for ASR in Indian languages [11], [12], [13], [14], [15], [16], [17], [18], [19]. Few speech corpora are available containing English utterances spoken by Indian speakers [20], [21], [22]. However, those do not meet the requirements of ASR. DA-IICT corpus was collected from 137 speakers for speaker recognition task [20]. In this corpus, only 10 mins of recording was collected from each speaker, further, of which only 20% of data is available with sentence transcriptions. Thus, the data from this corpus is limited for ASR task. A speech database collected from KIIT containing recordings from 100 speakers each speaking a set of 630 sentences [21]. However, this data does not cover most of the major dialects in India as well as the data from only one speaker is processed, thereby limiting its usage for the ASR task. IITKGP-MLILSC corpus was collected for language identification task and it contains only 82 mins of English utterances spoken by 25 speakers [22].

Apart from the corpora collected within India, there exist a few corpora, collected outside India. CSLU telephone speech corpus was collected from Hindi and Tamil speakers [23]. L2-ARCTIC corpus consists of one hour of speech data from one male and one female Hindi speakers [24]. Both databases are limited in data size as well as in number of dialects. Among all the existing corpora, L2-ARCTIC contains manually annotated phoneme transcriptions, which could be useful for the task of mispronunciation detection and diagnosis. However, only 300 utterances were annotated among all recordings of both the Hindi speakers. Apart from the above corpora, English speech data was collected by Indian Government organizations such as Linguistic Data Consortium for Indian Languages (LDC-IL) [25] and Technology Development for Indian Languages (TDIL) [26]. However, these data are also limited in dialects

978-1-7281-2449-0/19/$31.00 ©2019 IEEE

and data size.

In India, there is a lot of demand for English language learning since it is a major language of communication in administration, law and education [27]. Also, most Indians pursue their career abroad [28], where English is a medium of communication. In order to analyse the language proficiency of Indian learners, it is required to adapt the techniques of computer assisted language learning (CALL) to Indian nativity. In CALL, ASR and mispronunciation detection and diagnosis are important components. In the past, the ASR models for these applications were designed with either native English speakers data [29], [30] or data belonging to a small set of Indian native speakers belonging to fewer accents [31]. On the other hand, in the applications of mispronunciation detection and diagnosis, analysis were confined to a small set of pronunciation errors made by Indian speakers belonging to fewer nativities [32].

In order to facilitate the analysis for many Indian native speakers in the applications of CALL, we obtain recordings of 2342 phonetically rich English stimuli from 80 Indian speakers. The stimuli are taken from TIMIT corpus [33]. We refer the collected corpora as Indic TIMIT. A subset of 2342 recordings is manually annotated to obtain phoneme transcriptions from two linguists to reflect their pronunciation, for which, the recordings are randomly chosen from the entire corpora ensuring one recording per stimuli. Further, to capture Indian specific pronunciation variabilities, we construct a lexicon, referred to as Indic English Lexicon, applying Indian specific mispronunciations while speaking English on to a native English lexicon deduced from CMU [34] and Beep [35] lexicons. We conduct the experiments to know the effectiveness of the Indic TIMIT and Indic English Lexicon considering word error rate (WER) and phoneme error rate (PER) as objective measures respectively.

II. Recording of Indic TIMIT

A. Subjects

India is known for its language diversity, it has more than 1652 dialects/languages [36], [37] out of which 22 are scheduled languages (based on 2001 census of India) [38]. It is impractical to record voice from the subjects belonging to all 1652 dialects/languages separately. Instead, we consider languages which are scheduled languages and spoken by majority of the population. These languages share similar properties when those are demographically close. Further, these language properties are influenced by the language family from which those are originated as well as by the language families of other closely related languages. Based on demographic differences, these languages are divided into six regions – 1) North East, 2) East, 3) North, 4) Central, 5) West and 6) South. The languages within these six regions are influenced mostly by the following four family of languages [39] – 1) Indo-Aryan, 2) Dravidian and 3) Austro-Asiatic 4) Tibeto-Burman. Table I shows the languages (scheduled) considered for Indic TIMIT from each region, the percentage of the population that speaks the respective language and the number of subjects considered for the recording. The table also shows the language family from which each of the considered language is originated [40] and the language families (blue colored text) influencing that language. It is to be noted that ∼90% of the Indian population speak these considered language together. Thus, we believe that it is sufficient to consider the subjects from these native languages in order to cover accent variabilities in most of the Indian population.

TABLE I
MAJORITY OF THE LANGUAGES SPOKEN IN EACH OF THE SIX REGIONS AS WELL AS GROUPING OF THE LANGUAGES CONSIDERED.

Region	Native language	Population percentage	Originated and/or influenced language family	Number of subjects (M/F) recorded	Grouping
North East	Assamese	1.28	Indo-Aryan Austro-Asiatic [41] Tibeto-Burman [41]	2 (0/)	Group-1
	Nepali	0.28	Indo-Aryan Tibeto-Burman [42]	1 (0/1)	
	Manipuri	0.14	Tibeto-Burman	1 (1/0)	
East	Bengali	8.10	Indo-Aryan Austro-Asiatic [43]	8 (4/4)	
	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 [44]	8 (3/5)	Group-4
	Telugu	7.19	Dravidian Indo-Aryan [44]	8 (5/3)	
	Malayalam	3.21	Dravidian	8 (3/5)	Group-5
	Tamil	5.91	Dravidian	8 (5/3)	

From the table, it is observed that the languages within each region are not confined to same language family as well as the population that speaks the respective languages varies across the regions. It is also observed that the same language family influences many languages in multiple regions. Considering these, we group the languages of from all six regions into five groups as shown in the table to obtain similar properties within a group by maintaining discrimination across the groups. In the first group, we consider all the languages belonging to both North East and East regions due to the influence of Austro-Asiatic language family in both the regions. Also, Bengali is spoken in both these regions as well as it has influences on Oriya and Maithili. The languages in North and central regions are grouped together due to commonality of Hindi spoken population. The West region is considered as the third group. However, we divide the south region into two groups, where one group contains Kannada and Telugu, which have influences from both Dravidian and Indo-Aryan language families and other group contains Tamil and Malayalam that mostly do not have influences from other language families.

For the recordings, we consider a total of 16 subjects from each of these groups with equal male to female ratio to maintain uniformity across the groups. The age of the subjects vary from 18 to 60 years with an average age of 25.42 years with standard deviation of 6.05 years. The subjects are mostly undergraduate or post graduate students

from Indian Institute of Science, Bangalore, India, while a few are working staff in the same institute. Prior to recording, we obtain consent from every subject as recommended by the institute ethics committee. On completion of recording, we provide remuneration as token of appreciation to each subject. We observe that the subjects have variabilities in their pronunciation ability of reading English.

B. Recording setup

We collect recordings from all 80 subjects when each of them read a given stimulus. Figure 1 shows the setup considered for the recording. It contains a display device for showing stimuli and a recording device. The entire recording process is performed with the help of an operator, who, during recording, checks if there is any error (insertion, deletion of substitution of words) in speaking a given sentence. Figure 1 also shows an exemplary screen-shot of the display, where a sentence to read is shown along with two control buttons. When a stimulus (sentence) is clicked, the green box enclosing the sentence turns into red and signals the subject to start speaking. Once the sentence is completely read, on click, the red box turns to cyan to indicate that the recording is completed. During the recording, the operator carefully listen to the subject's speech to spot error if there is any. If any error occurs, the subject is asked to read the sentence once again. On successful completion of a sentence's recording, the next/previous sentence is displayed with the click of Next/Previous button.

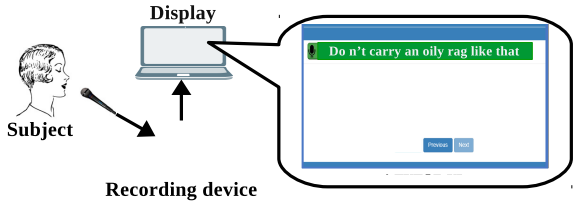

Fig. 1. Proposed setup for the recording

In the recording, we consider all 2342 unique sentences taken from TIMIT corpus. We divided all the stimuli into 16 parts, among which the first to fifteenth parts contain 150 stimuli and the sixteenth part contains 92 stimuli. The subject is allowed to take a break after completion of each part for any duration that the subject wants to. However, during the recording of each part, the subject is not allowed to take a break. The recordings of all the parts from each subject are done in multiple sessions. We recorded all the stimuli in a quiet room. For display, we consider Lenovo think pad edge laptop containing 2GB RAM, Windows-8 operating system with model no: E580. We consider Zoom H6 mixer along with Rode procaster microphone in the recording. We recorded all the stimuli in each part as a single recorded file, following which, manual segmentation of the entire recording is done to obtain the audio for each stimuli. All the recordings are done in 48000Hz sampling rate with 16bit PCM format. The audio of each sentence is down-sampled to 16000Hz and named with following convention – "LANG_GEN-ID_AGE_STIMULIno.wav" , where 'LANG' indicates the native language of a subject, 'GEN' takes two values M (male) or F (female), 'ID' is a number indicating the subject identity 'AGE' is two digit value indicating their age and 'STIMULIno' is the number in the stimuli list considered for the recording.

III. Annotation of Indic TIMIT

A. Manual annotation of phoneme transcriptions

A subset of recordings is selected from the entire recordings for the manual annotation. The manual annotation is performed through an online-interface with help of two linguists. One annotator has a PhD degree in linguistics and is working in central institute of Indian languages (CIIL), Mysore, India. The other annotator has a MSc degree in linguistics and is working in CIIL, Mysore, India. The stimuli for annotation are selected such that it covers the entire set of 2342 unique sentences in the TIMIT maintaining one recording per sentence. Out of 2342 selected stimuli, a total of 512, 512, 295, 511 and 512 stimuli are considered randomly from the speakers belonging to first, second, third, fourth and fifth groups respectively.

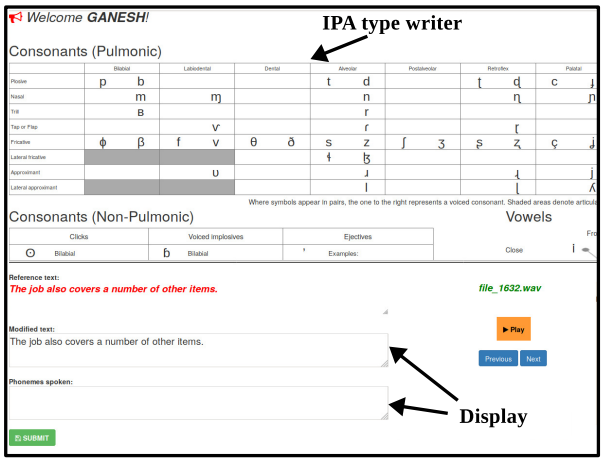

Fig. 2. A screen-shot of the online interface used in the manual annotation process. In the screen-shot, a box displaying instructions is cropped for better visibility of the remaining.

1) Online interface: We collect the phoneme transcriptions in an utterance using an online interface a screen-shot of which is shown in Figure 2. It has two main parts – IPA type writer and Display. Using the IPA type writer, the annotator can enter phoneme symbols in IPA format and the same will be shown in the box titled "Phoneme spoken" at the bottom of the display. Above this box, a box for modifying the text is provided. The interface also provides the following four buttons for navigation – 1) Play, 2) Previous, 3) Next, and 4) Submit. Further, it displays the file that is currently being annotated (green colored text) and its respective sentence transcriptions (red colored text). On click of the 'Play' button, the audio of the current sentence is played. The 'Previous' and 'Next' buttons are used to select the previous and next stimuli respectively. In the interface, instructions are provided for the annotation and the annotator is asked to provide phoneme transcriptions accordingly. In case of any violation of the instructions, the interface does not allow to submit and it displays an error message when the 'Submit' button is clicked.

If there is no error, the annotator can submit and head for the following stimuli. In the annotation, the instructions are provided to ensure that the word boundaries are marked by '∼' in the phoneme transcription. The instructions (which are cropped from the shown screen-shot in Figure 2 for better visibility of the remaining) are detailed as follows:

- 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). See the exemplary transcriptions for the text "I didn't hurt you" for the uttered phonemes in the recording "i d ɪ d n t h ɝ tʃ u". In this example, there is co-articulation between the words "hurt" and "you".
- Correct: in the "Modified text" and "Phoneme spoken" boxes, the entries should be "I didn't hurt-you", "i∼d ɪ d n t∼h ɝ tʃ u" but not "I didn't hur t-you", "i∼d ɪ d n t∼h ɝ∼tʃ u".

2) Analysis on annotated transcriptions: In order to know the pronunciation variations by Indian speakers, we compare the phoneme transcriptions from the annotator with respect to those available in TIMIT. However, the phoneme set used by the annotators and that available in TIMIT are different and also the former is in IPA and the later is in ARPAbet format. In order to obtain identical phoneme set in both the phoneme transcriptions, we collect the mappings between the IPA and ARPAbet symbols from the annotators. Following this, we map all the phoneme symbols in the corpora to a set having 40 phonemes used in CMU pronunciation dictionary [34]. In Indic TIMIT, we provide files containing original phoneme transcriptions, as well as in ARPAbet format and in ARPAbet mapped with 40 phoneme set. We also provide the files containing IPA to ARPAbet symbol mapping and from that to CMU phoneme set mapping. Considering the mapped phoneme transcriptions, we perform string alignment [45] using phoneme transcriptions available in the TIMIT corpora for all 2342 stimuli. However, in TIMIT, for a given stimuli there exist multiple recordings. Among these, we use the one whose phoneme transcription results in the least alignment distance. Based on this string alignment, Table II shows the percentage of correctly uttered phonemes and phoneme errors (insertions, deletions and substitutions) made by the Indian speakers. From the table, it is observed that the percentage of correct and erroneous phonemes are comparable. Thus, pronunciation of Indian speakers is largely different from that of the native English speakers.

TABLE II
PERCENTAGE OF CORRECTLY UTTERED PHONEMES AND PHONEME ERRORS

Corrects	Insertions	Deletions	Substitutions
51.77	14.64	0.71	32.88

IV. INDIC ENGLISH LEXICON

A. Indian speaker specific pronunciation variations

In general, the pronunciation of the Indian speakers differs from native English pronunciation due to the pronunciation errors. The typical variations made by the Indian speakers are listed in Table III, which are collected from the work by Sailaja [46]. From the table, it is observed that the pronunciation variations depend on the contexts based on the phonemes in the native English pronunciations and/or based on the letters in English words. Considering these rules, we obtain new word pronunciations by incorporating the variations to the existing word pronunciations in a native English lexicon. Following this, we propose to obtain an Indian speaker specific lexicon, referred to as IndicLexicon, by augmenting a native English lexicon with these new entries. We hypothesize that IndicLexicon could be useful to detect pronunciations made by the Indian learners in an automatic manner for the benefit of CALL applications. Further, it could be useful in the applications of ASR as well.

TABLE III
TYPICAL PRONUNCIATION VARIATIONS MADE BY THE INDIAN SPEAKERS WHILE SPEAKING ENGLISH

Phoneme specific context rules			
Previous phoneme	Target phoneme	Next phoneme	Indian specific variations
Vowel	Plosive	Vowel	Plosive is voiced
Nasal	Plosive	Any	Plosive is voiced
Any	Dipthong (except /aɪ/, /aʊ/)	Any	Substituted with long vowels
Any	/ʒ/	Any	Substituted with dʒ
Any/None	/θ/	Any	Substituted with /t̪ʰ/ or /t̪/
Any/None	/ð/	Any	Substituted with with /d̪/
None	Front vowel	Any	Phoneme /j/ is inserted before the vowel
None	Back vowel	Any	Phoneme /w/ is inserted before the vowel
None	/w/	Any	Phoneme /w/ is deleted
Any	/tʃ, dʒ, s, z, ʃ, ʒ/	Any	Substituted with /ɛs/ or /ɛz/ or /əz/
Any	/f/	Any	Substituted with /pʰ/
Any	/v,w/	Any	Substituted with /bʰ/
None	consonant	consonant	vowel is inserted before or within both the consonants
Letter specific context rules			
Previous letter	Target Letter	Next letter	Indian specific variations
Any	r	Any consonant	Phoneme /r/ is produced
Any	s	t	Phoneme /ʃ/ or /s/ is produced
Any	n	g	Both /ŋ/ and /g/ are produced
Any	r	None	Phoneme /r/ is produced
Both letter and phoneme dependent context rules			
Previous letter	Target letter	Next phoneme	Indian specific variations
Any	Double consonants	Short vowel	Geminate articulation

B. Lexicon construction

For given native English word pronunciation, we modify it by incorporating Indian speaker specific variations at the locations where the context criteria are met. In order to check the context criteria in the phoneme specific rules, it is sufficient to use the phoneme sequence in the word pronunciation. However, for the remaining two sets of context criteria, it is required to consider both the letters in the word and the phoneme sequence. This is because in order to incorporate India speaker specific variation, it is necessary to know a location where the rule to be placed. The location can be

obtained by knowing the mapping between the letters and the phonemes in the word pronunciation. But, in general, there is no one-to-one mapping between letters and phonemes for a given pronunciation [47]. As an example, for an eight letter word "Abnormal" one of the pronunciation in a native English lexicon contains only six phonemes ('ə b n ɔ: m l'). In order to circumvent this, we propose to consider letter-to-phoneme aligner [47] and obtain modified pronunciation following the three steps depicted in the block diagram shown in Figure 3. In the first step, for a given word, we map the letters in the word to the phonemes in the word pronunciation using letter-to-phoneme aligner. In the second step, we select a sub-set of Indian specific rules from the look-up table (Table III) by checking the context criteria in the table considering letter and phoneme mappings. In the third step, we obtain modified pronunciations using all k-combinations of rules in the sub-set, where k varies from 1 to the sub-set size. For each combination, the modified pronunciation is obtained by incorporating rules at the locations where the respective context criteria are met.

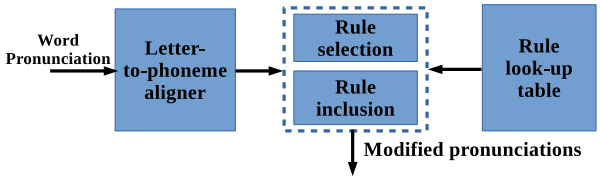

Fig. 3. Block diagram describing the steps involved in the creation of the Indic English lexicon

We obtain Indic English lexicon from a native English lexicon obtained by augmenting all the word pronunciations from CMU, Beep and the lexicon available in the TIMIT corpus. We use M2M aligner [47] for letter-to-phoneme alignment. A total of 3,62,751 entries are observed in the native English lexicon, which results into a total of 5,42,917 entries in Indic English lexicon.

V. Efficacy of Indic TIMIT and Indic Englich Lexicon

As a preliminary study, we analyze the benefit of the Indic TIMIT and Indic English Lexicon. For this, we perform the experiments in an ASR framework. The first one is analysed based on ASR performance in terms word error rate (WER). The second one is based on phoneme error rate (PER) considering forced-alignment process.

A. ASR with Indic TIMIT data

Experimental setup: We build the ASR models using Kaldi speech recognition tool-kit [48]. In the ASR modeling, we consider TDNN (time-distributed neural network) implementation from the tool-kit. For comparison, we train ASR models with TIMIT data belonging to the train set. Further, in order to build ASR model with Indic TIMIT data, we divide it into train and test sets. The train set contains all 1636 stimuli in TIMIT train set from randomly chosen 63 speakers maintaining region and gender balance. The test set contains the remaining 706 stimuli from the remaining speakers. We build the language model from the sentences in the training set. We consider the native English lexicon used in obtaining Indic English lexicon.

Results: Table IVa shows the WER obtained on the test set considering the models trained with both TIMIT and Indic TIMIT data. From the table, it is observed that the WER are lower when the model is trained with Indic TIMIT data compared to that with TIMIT data. The large benefit in the WER could be because of Indic TIMIT data size is larger than the TIMIT data size and mismatched acoustic characteristics between Indian and native English speakers. This indicates the effectiveness of Indic TIMIT data.

TABLE IV
Analysis of benefit of Indic TIMIT and Indic English lexicon considering performances of a) ASR and b) Forced-alignment.

(a)

ASR performance		
	TIMIT	Indic TIMIT
WER	93.41	15.02

(b)

Forced-alignment performance		
	Native lexicon	Indic lexicon
PER	32.49	28.79

B. Forced-alignment with Indic English lexicon

Experimental setup: We train the models using train set from Indic TIMIT considering native and Indic English lexicon separately. Considering these models, we obtain phoneme transcriptions by performing forced-alignment on the sub-set of 706 recordings from the test set for which manual phoneme transcriptions are available.

Results: Table IVb shows the PER obtained on the sub-set of the test set using native and Indic English lexicon. From the table, it is observed that the PER is lower when Indic English Lexicon is used compared to that with native English lexicon. This shows the benefit of Indic English Lexicon. This could be because the Indic English lexicon contains erroneous pronunciations from Indian learners, which, in turn, helps in achieving lower phoneme error rate on Indian learners data.

VI. Conclusions

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 and it is made available with the corpus. The corpus contains ∼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. Though the phoneme transcriptions are provided in the Indic TIMIT corpus for one set covering all 2342 stimuli across all five regions, further works are required to annotate five sets, where each set contains all 2342 stimuli from each region considering uniform number of stimuli per speaker and multiple annotators.

References

[1] B. Council, "The English effect: The impact of English, what its worth to the UK and why it matters to the world," *London: British Council*, 2013.

[2] M. Mehrabani, J. Tepperman, and E. Nava, "Nativeness classification with suprasegmental features on the accent group level." *Interspeech*, pp. 2073–2076, 2012.

[3] M. Swan, "The influence of the mother tongue on second language vocabulary acquisition and use," *Vocabulary: Description, acquisition and pedagogy*, pp. 156–180, 1997.

[4] D. Van Compernolle, "Recognizing speech of goats, wolves, sheep and non-natives," *Speech Communication*, vol. 35, no. 1, pp. 71–79, 2001.

[5] C. Hua, W. Qiufang, and L. Aijun, "A learner corpus-ESCCL," *Speech Prosody*, pp. 155–158, 2008.

[6] R. W. Ng, A. C. Kwan, T. Lee, and T. Hain, "Shefce: A Cantonese-English bilingual speech corpus for pronunciation assessment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5825–5829, 2017.

[7] Y. Chen, J. Hu, and X. Zhang, "Sell-corpus: an open source multiple accented Chinese-English speech corpus for l2 English learning assessment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7425–7429, 2019.

[8] Q. Wen, L. Wang, and M. Liang, "Spoken and written English corpus of Chinese learners," *Foreign Language Teaching and Research Press*, 2005.

[9] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken english," *Proceedings of Language Resources and Evaluation Conference (LREC)*, vol. 2, pp. 957–964, 2000.

[10] E. Izumi, K. Uchimoto, and H. Isahara, "The NICT JLE Corpus: Exploiting the language learners speech database for research and education," *International Journal of the Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, 2004.

[11] W. Lalhminghlui, R. Das, P. Sarmah, and S. Vijaya, "A Mizo speech database for automatic speech recognition," *International conference on speech database and assessments (Oriental COCOSDA)*, 2017.

[12] P. P. Shrishrimal, R. R. Deshmukh, and V. B. Waghmare, "Indian language speech database: A review," *International journal of Computer applications*, vol. 47, no. 5, pp. 17–21, 2012.

[13] C. Kurian, "A review on speech corpus development for automatic speech recognition in Indian languages," *International Journal of Advanced Networking and Applications*, vol. 6, no. 6, p. 2556, 2015.

[14] B. Das, S. Mandal, and P. Mitra, "Bengali speech corpus for continuous auutomatic speech recognition system," *International conference on speech database and assessments (Oriental COCOSDA)*, pp. 51–55, 2011.

[15] M. Shridhara, B. K. Banahatti, L. Narthan, V. Karjigi, and R. Kumaraswamy, "Development of Kannada speech corpus for prosodically guided phonetic search engine," *International conference on speech database and assessments (Oriental COCOSDA)*, pp. 1–6, 2013.

[16] K. Samudravijaya, P. Rao, and S. Agrawal, "Hindi speech database," *Sixth International Conference on Spoken Language Processing*, 2000.

[17] K. Prahallad, E. N. Kumar, V. Keri, S. Rajendran, and A. W. Black, "The IIIT-H indic speech databases," *Proceedings of Interspeech*, pp. 2546 – 2549., 2012.

[18] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "IITKGP-SESC: speech database for emotion analysis," *International conference on contemporary computing*, pp. 485–492, 2009.

[19] T. Godambe, N. Bondale, K. Samudravijaya, and P. Rao, "Multi-speaker, narrowband, continuous Marathi speech database," *International conference on speech database and assessments (Oriental COCOSDA)*, pp. 1–6, 2013.

[20] H. A. Patil, S. Sitaram, and E. Sharma, "DA-IICT cross-lingual and multilingual corpora for speaker recognition," *Seventh International Conference on Advances in Pattern Recognition*, pp. 187–190, 2009.

[21] S. S. Agrawal, S. Sinha, P. Singh, and J. Ø. Olsen, "Development of text and speech database for Hindi and Indian English specific to mobile communication environment." *Language resources and evaluation conference (LREC)*, pp. 3415–3421, 2012.

[22] S. Maity, A. K. Vuppala, K. S. Rao, and D. Nandi, "IITKGP-MLILSC speech database for language identification," *National Conference on Communications (NCC)*, pp. 1–5, 2012.

[23] R. A. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," *Fourth European Conference on Speech Communication and Technology*, 1995.

[24] G. Zhao, S. Sonsaat, A. O. Silpachai, I. Lucic, E. Chukharev-Khudilaynen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: a non-native english speech corpus," *Perception Sensing Instrumentation Lab*, 2018.

[25] "LDC-IL: Linguistic Data Consortium for Indian Languages," *Online. http://www.ldcil.org/*.

[26] "TDIL: Technology Development for Indian Languages," *Online. http://tdil.mit.gov.in/*.

[27] A. Dey and P. Fung, "A Hindi-English code-switching corpus." *Language resources and evaluation conference (LREC)*, pp. 2410–2413, 2014.

[28] "TOEFL: Test of English as a Foreign Language," *URL http://www.ets.org/toefl*.

[29] S. Joshi and P. Rao, "Acoustic models for pronunciation assessment of vowels of Indian english," *2013 International Conference Oriental COCOSDA*, pp. 1–6, 2013.

[30] C. Bhat, K. Srinivas, and P. Rao, "Pronunciation scoring for Indian english learners using a phone recognition system," *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia*, pp. 135–139, 2010.

[31] A. Garud, A. Bang, and S. Joshi, "Development of hmm based automatic speech recognition system for Indian english," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–6, 2018.

[32] V. V. Patil and P. Rao, "Detection of phonemic aspiration for spoken Hindi pronunciation evaluation," *Journal of Phonetics*, vol. 54, pp. 202–221, 2016.

[33] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[34] R. Weide, "The CMU pronunciation dictionary, release 0.6," *Carnegie Mellon University*, 1998.

[35] A. Robinson, "BEEP pronunciation dictionary," *Retrieved from World Wide Web: ftp://svr-ftp. eng. cam. ac. uk/pub/comp. speech/dictionaries/beep. tar. gz*, 1996.

[36] A. H. Unnibhavi and D. Jangamshetti, "Development of Kannada speech corpus for continuous speech recognition," *International Journal of Computer Applications*, vol. 975, p. 8887.

[37] "Read on to know more about Indian languages," *URL: https://mhrd.gov.in/sites/upload_files/mhrd/files/upload_document /languagebr.pdf, last accessed on 20-06-2019*, 2001.

[38] "Office of the Registrar General & Census Commissioner India, Part A: Distribution of the 22 scheduled languages - India, States & union Territories - 2001 Census," *URL: http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online /Language/parta.htm, last accessed on 12-06-2018*, 2001.

[39] J. Heitzman and R. L. Worden, *India: A country study*. Federal Research Division, 1995.

[40] "Office of the Registrar General & Census Commissioner India, Part A: Family-wise grouping of the 122 scheduled and non-scheduled languages–2001," *URL: http://censusindia.gov.in/Census_Data_2001/Census_Data_Online /Language/statement9.aspx*, 2001.

[41] D. Moral, "North-East India as a linguistic area," *Monkhmer Studies*, pp. 43–54, 1997.

[42] B. H. Hodgson, *Essays on the languages, literature, and religion of Nepal and Tibet: together with further papers on the geography, ethnology, and commerce of those countries*. Trübner & Company, 1874.

[43] P. Sidwell, "The Austroasiatic central riverine hypothesis," *Journal of Language Relationship*, no. 4, pp. 117–134, 2010.

[44] S. Sridhar, "Linguistic convergence: Indo-Aryanization of Dravidian languages," *Lingua*, vol. 53, no. 2-3, pp. 199–220, 1981.

[45] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of the ACM (JACM)*, vol. 21, no. 1, pp. 168–173, 1974.

[46] P. Sailaja, *Dialects of English: Indian English*. Edinburgh University Press, 2009.

[47] S. Jiampojamarn, G. Kondrak, and T. Sherif, "Applying many-to-many alignments and hidden Markov models to letter-to-phoneme conversion," *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 372–379, 2007.

[48] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," *IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.