voisTUTOR corpus: A speech corpus of Indian L2 English learners for pronunciation assessment.

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

Aparna Srinivasan
Electrical Engineering
Indian Institute of Science
Bengaluru, India
aparnasrinivasa@iisc.ac.in

Chandana Srinivasa
Computer Engineering
Arizona State University
Tempe, United States of America
csriniv3@asu.edu

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

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

***Abstract*—This paper describes the voisTUTOR corpus, a pronunciation assessment corpus of Indian second language (L2) learners learning English. This corpus consists of 26529 utterances approximately totalling to 14 hours. The recorded data was collected from 16 Indian L2 learners who are from six native languages, namely, Kannada, Telugu, Tamil, Malayalam, Hindi and Gujarati. A total of 1676 unique stimuli were considered for the recording. The stimuli were designed such that they ranged from single word stimuli to multiple word stimuli containing simple, complex and compound sentences. The corpus also consists of ratings representing overall quality on a scale of 0 to 10 for every utterance. In addition to the overall rating, unlike the existing corpora, a binary decision (0 or 1) is provided indicating the quality of the following seven factors, on which overall pronunciation typically depends, – 1) intelligibility, 2) phoneme quality, 3) phoneme mispronunciation, 4) syllable stress quality, 5) intonation quality, 6) correctness of pauses and 7) mother tongue influence. A spoken English expert provides the ratings and binary decisions for all the utterances. Furthermore, the corpus also consists of recordings of all the stimuli obtained from a male and a female spoken English expert. Considering factor dependent binary decisions and spoken English experts' recordings, voisTUTOR corpus is unique compared to the existing corpora. To the best of our knowledge, there exists no such corpus for pronunciation assessment in Indian nativity.**

***Index Terms*—non-native English corpus, Indian speakers, overall rating, factor-specific binary decision**

I. Introduction

English has become the global language and the most preferred language for communication between speakers of different nativities [1], [2]. Consequently, it has become important to learn English as second language (L2), as it would help to prevent miscommunication due to incorrect pronunciation. Especially in a multilingual nation like India, English is generally used as the language of communication in administration, law and education [3]. Furthermore, with the advent of globalization in India, effective English communication also helps to get lucrative job opportunities [4]. In general, most of the L2 learners improve their English pronunciation using the applications of Computer Assisted Pronunciation Training (CAPT) [5]. These applications provide a self-learning platform by performing an automatic pronunciation assessment on learner's utterance and by providing a feedback [6]. The pronunciation assessment is typically carried out using data-driven approaches thereby requiring large speech corpora from L2 learners [7]. However, such corpora are limited in Indian context.

There are a few corpora from L2 learners that are collected within India. For example, Chandel et al. collected 4860 utterances from a total of 243 call center candidates which were rated on scale of 1 to 4 for overall quality [8]. Apart from this corpus, there exist corpora that were collected outside India. However, the number of utterances from Indians are limited in them. Cheng et al. developed a corpus consisting of 3380 utterances that were rated on a scale of 0 to 6 for overall quality [9]. Similarly, the C-AuDiT corpus consists of utterances from native speakers of German, Spanish, Italian and Hindi that were rated on a scale of 1 to 5 [10]. However, this corpus consists of only 329 utterance from each of the two native Hindi speakers. The CSLU corpus consists of 4925 utterances from native speakers of 22 languages including Hindi and Tamil [11]. This corpus also has scores for overall quality on a four point scale.

One of the corpora collected from the TOEFL iBT test by ETS, consists of utterances from speakers of different native languages which are rated on an overall quality scale of 1 to 4 [12]. Gruhn et al. developed a corpus from speakers belonging to China, France, Germany, Indonesia and Japan, along with a rating for each utterance [13]. Similarly, Witt et al. developed a corpus consisting of words uttered by speakers whose native languages were Spanish, Italian, Japanese and Korean [14]. In this corpus, each utterance is provided with a quality rating on a 4 point scale. The ISLE corpus consists of utterances

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

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

from German and Italian native speakers which were rated for overall quality on a scale of 1 to 5 [15]. Similarly, Neumeyer et al. developed the Japanese spoken English corpus, whose constituent utterances were provided a rating on a scale of 1 to 5 [16].

However, in all these corpora only one rating is provided that indicates the overall quality of an utterance. In contrast to these, there exist corpora that provide multiple ratings corresponding to the factors influencing the overall quality. The ERJ corpus consists of utterances from Japanese speakers who were rated on the quality of phoneme and intonation [17]. The CU-CHLOE corpus consists of utterances from Cantonese learners of English whose utterances were rated on scale of 1 to 4 indicating the degree of phoneme mispronunciation [18]. Imoto et al. developed the corpus from Japanese students and rated each syllable as primary stress, secondary stress or no stress [19]. The PF-STAR corpus consists of utterances from German school children in which each word is provided with a binary rating indicating mispronunciation [20]. In general, apart from the above factors (intelligibility, phoneme quality, phoneme mispronunciation, stress and intonation) considered, the overall quality of an utterance also depends on pause placement and mother tongue influence (MTI) [14], [21]–[30]. This indicates that excellent overall quality can be achieved only if the utterance complies with all dependent factors' requirements. In order to analyze the dependencies of these factors on the overall quality in the pronunciation assessment task, it is required to have a corpus that consists of ratings indicating overall quality as well as the quality of its dependent factors. To the best of our knowledge, there is no such corpus from either Indian or other native language L2 learners.

In order to cater these requirements in Indian context, we have collected a corpus under the voisTUTOR project [1] referred to as voisTUTOR corpus from Indian L2 learners. The corpus consists of 1676 utterances from 16 learners belonging to 6 native languages. The native languages considered were Kannada, Malayalam, Telugu, Tamil, Hindi and Gujarati. Each utterance was provided with binary decisions indicating the quality of the factors and a rating on a scale of 0 to 10 representing overall pronunciation quality. The factors considered were intelligibility, phoneme quality, phoneme mispronunciation, stress, intonation, pause placement and MTI. In addition, the corpus also consists of utterances belonging to the same set of stimuli recorded from a male and a female spoken English expert thereby, providing additional benefit in analysing a learner's pronunciation with respect to an expert. In total, the corpus consists of approximately 14 hours of recorded data. In this paper, we present a preliminary analysis to know how the overall pronunciation quality depends on these factors. Along with these, we also present variations of the ratings with respect to L2 learner's native language and utterance length in terms of word count.

II. Recording

We describe the recording process in the following four subsections – 1) stimuli, 2) L2 learner subjects, 3) spoken English expert subjects and 4) recording setup.

A. Stimuli

The stimuli used in the recording consist of words, phrases and sentences which were chosen from materials used for spoken English training [23], [31] and from the ISLE corpus [15]. Each stimulus was verified by an English teacher on any grammatical or spelling mistakes. After verification, the stimuli were grouped into four categories such that the complexity of the stimuli ranged from low to high. Further, in each category the stimuli were divided into multiple subcategories. A set of 1023 stimuli of minimal pairs in the first category were divided into 8 subcategories based on different phonological element in the pairs such as, fricatives, stops, nasals, glides & laterals, consonant sequences, vowels, diphthongs and vowel sequences. The number of stimuli in each subcategory was 283, 198, 30, 60, 200, 105, 60 and 87 respectively. The 113 stimuli in the second category were divided based on the following four types of intonation – glide up, glide down, dive and take off. Each subcategory in this category consists of 26, 32, 33 and 22 stimuli respectively. Similarly, the 189 stimuli in the third category were divided into single words, masked words, weak forms and phrases. The number of stimuli in each category were 48, 52, 51 and 38 respectively. Finally, the fourth category consists of 351 stimuli which were divided into subcategories depending on whether they were simple, complex, compound or long sentences, comprising of 90, 117, 100 and 44 stimuli respectively.

B. L2 learner subjects

Sixteen subjects were considered from two English training schools located in Bengaluru, India. The subjects were undergoing L2 English training at the time of recording. The native languages of the subjects include Malayalam, Kannada, Telugu, Tamil, Hindi and Gujarati. In each language, there are a total of 4 (3+1), 5 (1+4), 3 (2+1), 2(2+1), 1 (0+1) and 1 (0+1) subjects (male + female) respectively. In total, the 16 subjects comprising of 8 males and 8 females, were either undergraduate or postgraduate students in an age group of 19 to 25. These subjects were provided with a remuneration for the recordings.

C. Expert subjects

Two experts, one male and one female, of English language were identified in Bengaluru, India. The male expert was a voice over artist with over 20 years of experience and the female expert was a voice over artist and a spoken English teacher with over 25 years of experience. The male and female experts were remunerated for the recordings.

[1] This project has been funded by the Department of Science and Technology, Government of India, for developing an automated feedback system for degree of nativity in spoken English by Indian learners.

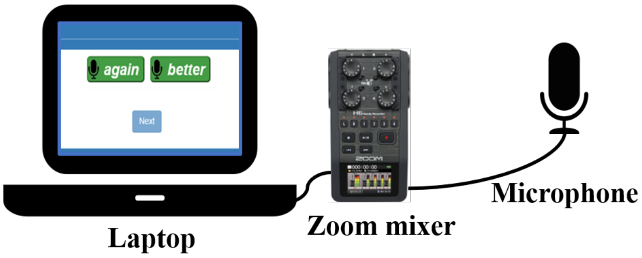


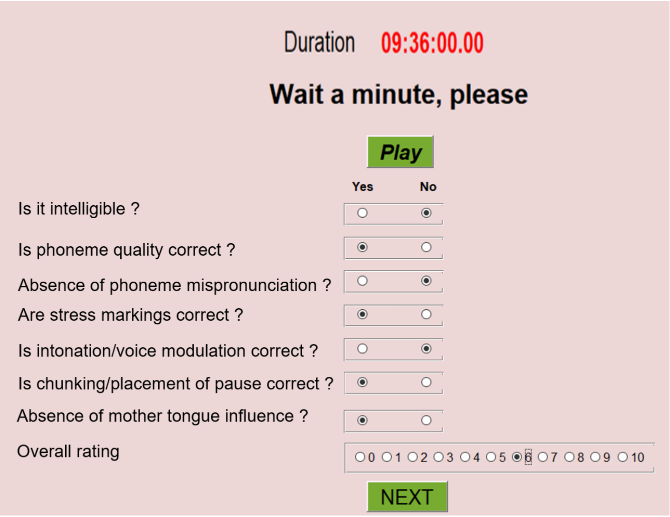
Fig. 1. The recording setup considered in collecting voisTUTOR corpus.

D. Setup

For recording of both learners and experts, the arrangements were made as shown in Figure 1, which was managed by a separate human operator. The figure shows an exemplary user interface (UI) that was created in order to facilitate the recording process. The UI displays every stimulus and a microphone symbol adjacent to each stimulus enclosed within a green box. Each stimulus can be recorded by clicking the corresponding microphone symbol. On click, the enclosing box turns red indicating that the recording is in progress. On click of the microphone symbol again, the recording is stopped and the enclosing box turns cyan indicating that the corresponding stimulus has been recorded. The UI also provides a "Next" button to navigate through the list of stimuli. While recording, the operator ensures that there is no word errors such as insertions, deletions and substitutions of the words. In case of any such errors, the learners were asked to repeat the respective stimuli. Figure 1 also shows the setup for the recording. A procaster microphone through a Zoom H6 mixer was connected to a laptop which was used for the UI. Both the learners and the experts were recorded using the procaster microphone, and the recordings were stored in the Zoom mixer. The subjects were recorded in a noise free environment at their respective English training centers whereas the experts were recorded at Indian Institute of Science, Bengaluru, India. The recordings took place in multiple sessions for covering all the stimuli recordings in a subcategory in a single recording. The recordings of all such sessions were manually segmented to obtain recording corresponding to each stimulus. Furthermore, the utterances containing any error were removed and such utterances were very few. In total, 13 hours 39 minutes of recording was done comprising of 26529 utterances from all 16 learners. The male and female experts had 1676 and 1668 utterances which resulted in a total of 58 minutes and 54 minutes of recordings respectively.

III. Binary decisions and ratings

In this process an expert provided two sets of integer values for each utterance: 1) binary decisions for the factors influencing the overall quality and 2) a rating in the range of 0-10 denoting the overall quality of the utterance. The factors considered were intelligibility, phoneme quality, pronunciation, stress, intonation, pause placement and MTI. Their respective binary decisions were collected through yes/no questions as shown in the exemplary scoring UI in Figure 2, with yes/no indicating a quality of 1/0 for the factors. Similarly, the overall rating was obtained in the range of 0-10 using the same UI, where 0 and 10 indicate poor and excellent overall quality respectively. The UI also displays the duration of recorded data that has been scored so far and the stimulus currently being scored. Furthermore, it provides a "Play" button which when clicked plays the utterance of the displayed stimulus from one of the subjects. The "Next" button needs to be clicked to submit the binary decisions and the overall rating. This also takes the expert to the next utterance.



Duration 09:36:00.00

Wait a minute, please

	Yes	No
Is it intelligible ?	○	●
Is phoneme quality correct ?	●	○
Absence of phoneme mispronunciation ?	○	●
Are stress markings correct ?	●	○
Is intonation/voice modulation correct ?	○	●
Is chunking/placement of pause correct ?	●	○
Absence of mother tongue influence ?	●	○

Overall rating ○0 ○1 ○2 ○3 ○4 ○5 ●6 ○7 ○8 ○9 ○10

Fig. 2. An exemplary screen-shot illustrating the UI used in providing the binary decisions for the factors and overall quality ratings.

All 1676 stimuli and their corresponding recordings from subjects were randomized for the scoring process. So, the stimuli appeared in a random order in the UI and once a stimuli was displayed, all the corresponding utterances were randomized and scored sequentially. Ratings were obtained from a female expert, a spoken English teacher, from whom the recordings were also collected. In total, it took about 177 hours to score all 26529 stimuli. In the scoring process, utterances corresponding to ∼2 hours of recording were randomly repeated to know the consistency of the expert providing the ratings and it was found that the the expert had consistency upto 70%.

IV. Statistics

In the following subsections we analyse the distribution of binary decisions given to each factor and the overall rating in terms of the length of the stimulus, the gender of the subjects, native languages and explore the relationship between the factors and overall quality.

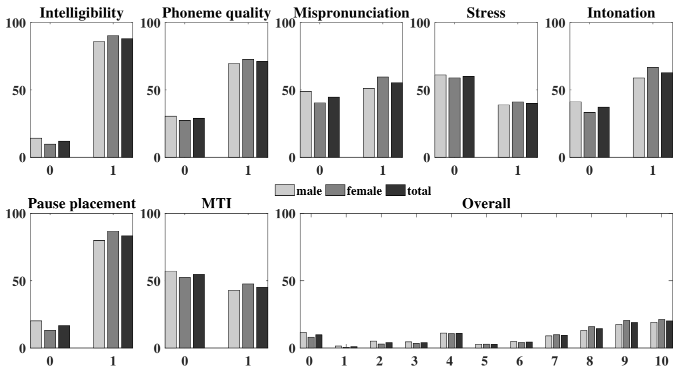


Fig. 3. Distribution of the binary decisions and the ratings of all 26529 utterances and all the utterances specific to male and female subjects.

A. Analysis of ratings and binary decisions and their variations

Figure 3 shows the percentage distribution of binary decisions for all factors and the percentage distribution of overall ratings, considering all utterances and utterances specific to male and female subjects separately. Considering the distribution for the factors, it is observed that for intelligibility, almost 88.10% of the utterances obtained binary decision 1. Similarly considering phoneme quality, intonation and pause placement it is observed that 71.12%, 62.80% and 83.37% of the utterances obtained binary decision 1. This indicates that a majority of the utterances are intelligible and have correct phoneme quality, intonation and pause placement. In the case of mispronunciation and MTI, it is observed that the percentage of utterances with binary decision of 0 is 44.63% and 54.75% respectively. This indicates that the utterances are likely to have mispronunciation and presence of MTI as much as they are likely to not have mispronunciation and absence of MTI respectively. However, for stress, it is observed that 60.02% of the utterances have obtained binary decision of 0 indicating that the subjects found it challenging to stress at the correct syllables. In general, for a majority of the factors except stress and MTI, more than 50% of the utterances obtained 1 as the binary decision thereby indicating that a majority of the utterances complied with each factor.

Considering the overall quality, it is observed that about 20.14% of utterances were given a rating of 10, which is also the rating obtained by the highest percentage of utterances. On the other hand, 9.77% of utterances were given a rating of 0. The rating of 9 has a lower percentage of utterances compared to the rating of 10 followed by the ratings 8 and 4. On the other hand, rating 1 corresponds to the least number of utterances. The distributions of each factor and overall rating for male and female have a trend similar to those of all speakers. This indicates that there no gender specific trend.

B. Analysis based on stimuli length and native language

Figure 4 shows the percentage distribution of binary decisions for all the factors and overall rating, in terms of utterance length and native language. The percentage distributions for each native language and each category of utterance length are stacked on top of one another respectively. For the distribution in terms of utterance length, the recordings were divided into three categories based on word count as, single words (word count is 1), short sentences (word count is in the range 2-5) and long sentences (word count greater than 5).

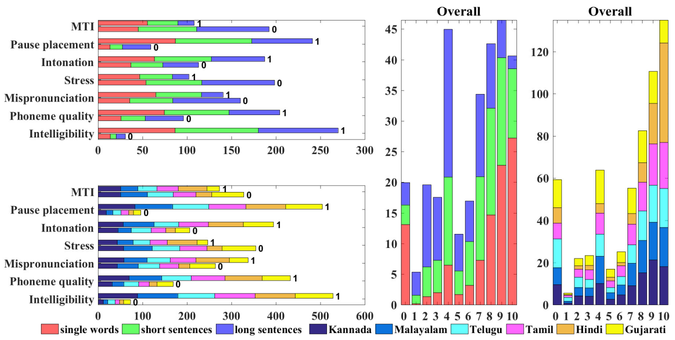


Fig. 4. Distribution of binary decisions and the ratings among the groups of utterances formed based on three sets of word-counts in an utterance and six native languages.

Considering the percentage distribution of binary decisions with respect to utterance length it is observed that, irrespective of utterance length, the majority of binary decisions were 1 for several factors including intelligibility, phoneme quality, intonation and pause. When the stimuli are short sentences it is observed that the majority of binary decisions were 0 for the factors mispronunciation and MTI. Considering long sentences, a similar distribution of binary decisions was observed for MTI. This indicates that as the complexity of the stimuli or the utterance length increases, the utterances fail to comply with factors. It is also observed that, most of the single word stimuli were scored with an overall rating of 9 and 10. Similarly, most of the short sentences were given an overall rating of 8 and 9. However, most of the long sentences were given an overall rating of 4. As observed in Figure 4, majority of the binary decisions are 0 for stress irrespective of stimuli length. Furthermore, for stress, it is also observed that as the complexity of the stimuli increases the difference between their corresponding binary score distributions increases. Additionally, the percentage of binary decisions of 0 for short sentences is relatively lower than that for long sentences. In general, it is observed that a higher complexity of a stimuli leads to a lower overall rating.

Considering the percentage distribution of binary decisions with respect to native languages it is observed that, for Kannada, Tamil, Telugu and Malayalam, most of the binary decisions are 1 for all factors except stress and MTI. Consequently, it is also observed that their overall rating is predominantly in the range of 8 to 10. This indicates that stress and MTI have an impact on the overall rating. Furthermore, the distribution of binary decisions with respect to these native languages is similar across all factors. For Gujarati, it is observed that most of the binary decisions are 0 for the factors

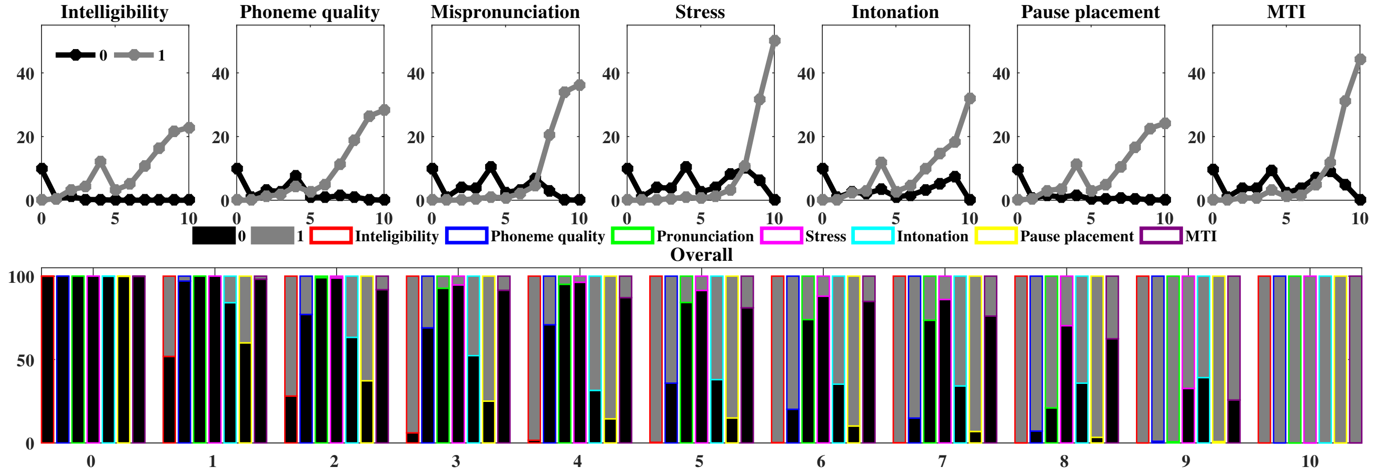


Fig. 5. Percentage of variations in binary decisions with respect to the overall ratings to explore the relation between overall quality and all of its seven dependent factors.

mispronunciation, stress and MTI. Consequently, it is also observed that the majority of overall rating obtained is 4. This indicates that mispronunciation has a greater impact on overall rating than stress and MTI. However, for the speaker whose native language is Hindi it is observed that most of the binary decisions are 1 for all factors and the majority of the overall rating obtained is 10. In general, from Figure 4 it is evident that the binary decisions made for each factor influences the overall rating.

C. Analysis of the relation between ratings and binary decisions

Figure 5 shows the percentage distribution of binary decisions with respect to overall rating for each factor using plots and bar graphs. Considering the plots, it is observed that as the overall rating increases from 0 to 10, for each factor the percentage of utterances having binary decision 1 increases and that of decision 0 decreases. However, the increase/decrease is not monotonic for all the factors especially around overall rating ranging from 2 to 9. In the case of intelligibility, phoneme quality, intonation and pause placement, it is observed that the plot corresponding to binary decision 1 undergoes a dip from rating 4 to 5. Similarly, the plot corresponding to binary decision 0 rises from rating 1 to 4 in phoneme quality, 1 to 4 and 5 to 8 for stress and MTI, and 3 to 4 and 5 to 9 for intonation. This indicates that the strength in the influence of factors varies across the ratings.

Considering the bar graphs, it is observed that when the overall rating is 0, then the binary decisions of all the factors are also 0 and vice versa. This indicates that a rating of 0 (10) is provided when there is compliance (non-compliance) with the requirements of all the factors. Considering the ratings in the range of 1 to 9 it is observed that the percentage of binary decision 1 slowly increases. However, it is interesting to notice that the increase in the percentage of score 1 belonging to the pronunciation and stress factors is more gradual than the remaining factors. This indicates that these factors have a stronger influence on the overall rating than the rest.

D. Preliminary experiments

1) Relation between factor-specific score and overall rating: A deep neural network based classifier consisting of 2 layers with 16 units each was used to learn the relation between the binary decisions of all factors and the overall rating. For this, the binary decisions of all factors were divided into 10 folds and trained based on a 10 fold cross-validation setup, with 8 folds for train, 1 for validation and 1 for test. The overall ratings were considered as the ground truth. The average accuracy across 10 folds on the test set was found to be 86.38% indicating a correlation between the binary decisions on all factors and the overall rating.

2) Speaking rate distribution across native languages: The speaking rate of each utterance was measured in terms of syllables/second. In order to determine the number of syllables in an utterance, first, its phonetic transcription was obtained using the Kaldi Automatic Speech Recognition toolkit [32]. The number of syllables was then computed considering a dictionary that provided the grouping of phonemes into syllables. It was observed that, in general, considering all utterances, the speaking rate varied from 2 to 6 syllables/second. A majority of the utterances belonging to native languages Kannada, Malayalam and Hindi were found to have speaking rate around 4 syllables/second. On the other hand, for Telugu and Gujarati the majority of the utterances had speaking rate in the range 4 to 6 syllables/second. For Tamil, it was found to be in the range of 2 to 4 syllables/second. From this, it is observed that native language affects the speaking rate of the subjects.

V. Conclusion

We have developed a non-native English corpus of Indian speakers for the applications of computer assisted language learning. The corpus consists of recordings of 1676 stimuli from 16 subjects representing 6 Indian native languages. The subjects were L2 English learners at the time of the recording. Recordings from a male and a female spoken English expert are also a part of the corpus. Furthermore, each utterance in the corpus has been given an overall rating and a binary value representing the quality of each factor that influences the overall quality. In total, the corpus contains approximately 14 hours of recorded data, covering 26529 utterances. This corpus can be used to explore the relationship between the factors and the overall rating, mispronunciation detection and correction in applications of CALL and evaluation of non-native speech with respect to an expert's speech. Further developments are required to obtain ratings and binary decisions from multiple spoken English experts and to evaluate consistency across the raters.

Acknowledgment

We would like to acknowledge the Department of Science and Technology, Government of India for funding the project. We also extend our gratitude to the management of Bengaluru School of English and English Age International School for providing their students and premises for recording. Further, we appreciate the support extended by the spoken English experts for recording and scoring the utterances in the corpus.

References

[1] D. Crystal, *English as a global language*. Cambridge University Press, 2012.

[2] B. Seidlhofer, "English as a lingua franca," *ELT journal*, vol. 59, no. 4, pp. 339–341, 2005.

[3] A. Dey and P. Fung, "A Hindi-English code-switching corpus." in *LREC*, 2014, pp. 2410–2413.

[4] B. Mahanta and R. B. Sharma, *English Studies in India: Contemporary and Evolving Paradigms*. Springer, 2019.

[5] H. S. Mahdi and A. A. Al Khateeb, "The effectiveness of computer-assisted pronunciation training: A meta-analysis," *Review of Education*, 2019.

[6] C. Yarra, P. Anand, N. Kausthubha, and P. K. Ghosh, "SPIRE-SST: An Automatic Web-based Self-learning Tool for Syllable Stress Tutoring (SST) to the second language learners." in *Interspeech*, 2018, pp. 2390–2391.

[7] M. G. OBrien, T. M. Derwing, C. Cucchiarini, D. M. Hardison, H. Mixdorff, R. I. Thomson, H. Strik, J. M. Levis, M. J. Munro, J. A. Foote *et al.*, "Directions for the future of technology in pronunciation research and teaching," *Journal of Second Language Pronunciation*, vol. 4, no. 2, pp. 182–207, 2018.

[8] A. Chandel, A. Parate, M. Madathingal, H. Pant, N. Rajput, S. Ikbal, O. Deshmukh, and A. Verma, "Sensei: Spoken language assessment for call center agents," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 711–716.

[9] J. Cheng, N. Bojja, and X. Chen, "Automatic accent quantification of Indian speakers of English." in *INTERSPEECH*, 2013, pp. 2574–2578.

[10] F. Hönig, A. Batliner, and E. Nöth, "Automatic Assessment of Non-native Prosody – Annotation, Modelling and Evaluation," in *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 21–30.

[11] T. Lander, *Foreign Accent English Release 1.2*. Linguistic Data Consortium, 2007.

[12] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[13] R. Gruhn, T. Cincarek, and S. Nakamura, "A multi-accent non-native English database," in *ASJ*, 2004, pp. 195–196.

[14] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

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

[16] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech communication*, vol. 30, no. 2-3, pp. 83–93, 2000.

[17] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," in *Proc. ICA*, vol. 1, 2004, pp. 557–560.

[18] H. Wang, H. Meng, and X. Qian, "Predicting gradation of L2 English mispronunciations using ASR with extended recognition network," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–4.

[19] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, "Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system," in *Seventh International Conference on Spoken Language Processing*, 2002, pp. 749–752.

[20] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-natives first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[21] A. Raux and T. Kawahara, "Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning," in *Seventh International Conference on Spoken Language Processing*, 2002, pp. 737–740.

[22] M. J. Munro and T. M. Derwing, "The foundations of accent and intelligibility in pronunciation research," *Language Teaching*, vol. 44, no. 3, pp. 316–327, 2011.

[23] J. D. O'Connor, *Better English Pronunciation*. Cambridge University Press, 1980.

[24] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[25] Q. Shi, K. Li, S. Zhang, S. M. Chu, J. Xiao, and Z. Ou, "Spoken English assessment system for non-native speakers using acoustic and prosodic features," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[26] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.

[27] M. A. Peabody, "Methods for pronunciation assessment in computer aided language learning," Ph.D. dissertation, Massachusetts Institute of Technology, 2011.

[28] P. Trofimovich and W. Baker, "Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech," *Studies in second language acquisition*, vol. 28, no. 1, pp. 1-30, 2006.

[29] J. Geertzen, T. Alexopoulou, B. Post, and A. Korhonen, "Native language effects on pronunciation accuracy in L2 English."

[30] E. Bada, "Native language influence on the production of English sounds by Japanese learners," *The Reading Matrix*, vol. 1, no. 2, 2001.

[31] L. James and O. Smith, *Get rid of your accent: The English pronunciation and speech training manual*. Business And Technical Communication Services, 2007.

[32] D. Povey, A. Ghoshal, and G. Boulianne, "The Kaldi Speech Recognition Toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.