2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA) | 979-8-3503-9856-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/O-COCOSDA202257103.2022.9997873

VOISTUTOR 2.0: A SPEECH CORPUS WITH PHONETIC TRANSCRIPTION FOR PRONUNCIATION EVALUATION OF INDIAN L2 ENGLISH LEARNERS

Priyanshi Pal[1], *Chiranjeevi Yarra*[2], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science (IISc), Bengaluru 560012, India
[2]Speech Lab, Language Technologies Research Center (LTRC), IIIT, Hyderabad, 500032, India

ABSTRACT

In computer assisted pronunciation training (CAPT), robust automatic models are critical for pronunciation assessment and mispronunciation detection and diagnosis (MDD). In the modelling, besides the audio data of second language (L2) learners, CAPT requires manually annotated ratings of overall pronunciation quality, and the MDD uses manually annotated phonetic transcriptions. Though the pronunciation quality and the mispronunciation are interdependent, to the best of our knowledge, none of the existing corpora contains both ratings and phonetic transcriptions. This could be due to the cost involved in obtaining phonetic transcriptions. However, a corpus with both kinds of information could benefit the researchers to obtain robust models by exploring the interdependencies. For addressing this, we develop voisTUTOR 2.0 corpus considering the existing voisTUTOR corpus referred to as voisTUTOR 1.0. We obtain phonetic transcriptions manually from a linguist for the entire Indian L2 learners' English audio data (26529 utterances approximately 14 hours) in voisTUTOR 1.0 for which overall quality ratings and binary scores of factors influencing the pronunciation quality are available. A preliminary analysis of voisTUTOR 2.0 suggests that the phonetic errors correlated with the ratings and the binary scores indicating mispronunciations and phoneme quality.

Index Terms— L2-English corpus, Pronunciation training, Mispronunciation detection and diagnosis, Phonetic transcriptions

1. INTRODUCTION

In most of the computer assisted language learning (CALL) applications, English is considered as a second language (L2) and has been a very popular target among L2 learners [1]. Learning correct pronunciation plays a critical role in language acquisition besides vocabulary and sentence construction. This, in turn, creates the need for developing tools, methods, and corpora to train, evaluate and grade the L2 learners' pronunciation automatically. The L2 learners influence the pronunciation quality, either at phonetic or prosodic (such as stress, intonation, and phrasing) levels. The influence at both the levels could be due to the differences in phoneme inventory and phonology of L2 learners' native language (L1) with those of L2 language [2]. Between the two levels, phonetic level L1 influences are required to compensate in the early stage of L2 learning as it limits the intelligibility of the linguistic content, i.e. message in the learners' utterance. On the other hand, the prosodic level L1 influences limit the paralinguistic content, i.e. emotional state, meaning in the learners' utterance. The phoneme level L1 influences cause phoneme mispronunciation (insertion, deletion and substitution of the phonemes in the correct pronunciation) and phoneme sound quality variations (L2 learners produce the sounds of L2 phonemes similar to those in their L1) [2].

In the literature, it has been emphasized that the phonetic transcriptions of learners' utterances can be exploited to model CALL applications for assessing L2 pronunciation quality as well as providing qualitative feedback on their pronunciation. The feedback models include identification of goodness of uttered phonemes and mispronunciation detection and diagnosis. Both of these feedback models require recorded audio from L2 learners with manually annotated phonetic transcriptions reflecting their pronunciations [3]. On the other hand, the pronunciation assessment models require L2 learners' audio with manually annotated ratings indicating the overall pronunciation quality. Thus a corpus with both overall ratings and phonetic transcriptions could be useful in building robust models. Further, both the information are not only a rich resource for building the models but also provide a pathway to exploit the relationship between overall ratings and phonetic influences [4]. Other than CALL applications, both the information could be useful in developing pronunciation quality based automatic phoneme recognizer under accented speech conditions. However, to the best of our knowledge, there are no corpora consisting of both overall ratings and phonetic transcriptions, mainly collected from Indian L2 learners of English.

In the literature, most speech corpora are available with speech data and orthographic transcriptions. A few corpora contain either manually annotated overall ratings or phonetic transcriptions. However, a very few corpora consist of both the information, but those are limited to L2 other than English. In the Indian context, a few corpora have been developed for the benefit of CALL. Murthy et al. [5] have collected

Kannada speech data from the students of 18-25 years age group with only annotated overall ratings for evaluating Kannada language pronunciation. Basu et al. [6] have collected Indian English speech data with canonical phonetic transcriptions. However, these transcriptions may not reflect speakers' uttered pronunciation. Yarra et al. [7] have collected English speech data from Indian speakers with annotated overall ratings and binary decision scores of the seven factors influencing the overall quality.

Other than the Indian context, Franco et al. [8] have collected Spanish speech data from American English speakers with both the annotated overall ratings and the phonetic transcriptions for only a sub-part of the data containing 2550 utterances. Zhang et al. [9] have developed English speech data from Mandarin speakers annotated with three-level ratings. They provided canonical phonetic transcriptions of the speech data, which may not reflect speakers' uttered pronunciation. Neumeyer et al. [10] have considered speech data of American speakers speaking French with both the overall ratings and phonetic transcriptions. Chotimongkoi et al. [11] have developed PELECAN corpus that consists of English speech data from Thai speakers. This corpus is annotated with phonetic transcriptions and ratings for each phoneme correctness; however, no overall ratings are available. Besides the corpora consisting of overall ratings, there are the corpora with only phonetic transcriptions and English speech data. These corpora include L2-Arctic [12], collected from Hindi speakers along with four other non-Indian native language speakers, and SELL [13], collected from Chinese and Mandarin speakers.

To facilitate better models for CALL, in this work, we develop a spoken English corpus, named voisTUTOR 2.0, containing recordings from Indian L2 learners of English in which phonetic transcriptions reflect the learners' pronunciation and overall ratings. This corpus also consists of binary scores for each recording, indicating the quality of seven factors influencing the overall quality. voisTUTOR 2.0 derived from the existing voisTUTOR corpus, referred to as voisTUTOR 1.0, collected by Yarra et al. [7]. voisTUTOR 1.0 consists of $26,529$ recorded audios with overall ratings and binary scores for all the seven factors. For voisTUTOR 2.0, all the recorded audios are transcribed by a linguist to reflect the learners' pronunciation. The addition of phonetic transcriptions to the voisTUTOR 1.0 makes the voisTUTOR 2.0 a unique corpus for building CALL applications in the Indian context. We also perform a preliminary analysis considering annotated phonetic transcriptions, the overall ratings and the binary scores. In the analysis, we compute a set of five measures that indicate the errors made by L2 learners with respect to the canonical pronunciations from the UK accented English. We study the correlations of these measures with the quality ratings and the binary scores of all the seven factors. We found an interdependence of the errors with the overall ratings and with the binary scores of most of the factors, which suggests the benefit of the included phonetic transcriptions for developing robust CALL applications.

2. VOISTUTOR 1.0

voisTUTOR 1.0 [7] was developed for the pronunciation assessment of Indian L2 learners' spoken English skills. It consists of 14 hours of speech data of $26,529$ utterances. The L2 learners were considered to have the following 6 Indian Native languages: Kannada, Telugu, Tamil, Malayalam, Hindi and Gujarati. The total number of unique stimuli in voisTUTOR 1.0 is $1,676$ and were selected according to sentence complexity, ranging from a single word to multiple words forming simple, compound and complex sentences. The stimuli was chosen from materials used for spoken English and ISLE corpus [14, 15] [16] . Further, the chosen stimuli cover the following aspects of pronunciation at the phoneme and prosody level: 1) Phonological elements (such as Fricatives, stops, nasals, glides & laterals, consonant sequences, vowels, diphthongs and semi-vowels), 2) Types of Intonation (glide up, glide down, dive and take off), 3) Single words, Masked words, weak forms and phrases, and 4) Sentences (Simple, complex, compound or long).

The chosen L2 learners were gender-balanced (8 male and 8 female) and selected from English training schools in Bangalore, India. The learners were either graduate or undergraduate, with the age group of 19 to 25 at the time of recording. The learners were undergoing English language training in their respective schools while collecting the data. The data was collected under read speech conditions in a studio-quality environment using a procaster microphone with a Zoom H6 mixer and a user inference, which showed the required stimulus.

Each audio was annotated with an overall quality rating on a scale of 0 to 10, where 0 and 10 indicates the lowest and the highest quality, respectively. Also, a set of seven binary scores (0 or 1) were also obtained for each audio for the respective seven factors that are influencing the overall quality. The considered factors are: #1) intelligible (1) or not (0), #2) phoneme sound quality is good (1) or not (0), #3) phoneme mispronunciation is absent (1) or not (0), #4) syllable stress is proper (1) or not (0), #5) intonation is proper (1) or not (0), #6) Chunking is proper (1) or not (0) and #7) MTI is present (1) or not (0). All the audios were annotated by an expert who had teaching experience in spoken English for about 30 years. The annotated ratings had a good intra-rater agreement.

Due to the uniqueness of voisTUTOR 1.0 in the Indian context and its richness in terms of audio data and annotations for pronunciation assessment, we consider enhancing its richness further by adding annotated phonetic transcriptions. The enhanced voisTUTOR 1.0 is referred to as voisTUTOR 2.0.

3. PHONETIC ANNOTATION AND PROCESSING

We obtain the annotations with the help of a linguist working at SPIRE Lab, Electrical Engineering, Indian Institute of Science, India. The linguist completed her PhD in linguistics and has about 15 years of experience transcribing before and after the PhD. The linguist was allowed to choose the phonetic set that best suits the transcribing. The linguist provided the transcriptions considering the IPA symbols, and we found that the total number of unique phoneme symbols used in the transcriptions is 128 [1] The transcribing process took approximately a total of 150 hours. In this process, we evaluate the consistency of the transcriptions obtained from the linguist considering the intra-rater agreement on the transcriptions. For this, we chose a 5% of the total data randomly and obtained its transcriptions by combining it with the original data randomly. We consider Cohen's Kappa [17] score as the metric to identify the agreement quality. The Kappa score is found to be 0.79, which indicates a high intra-rater agreement. Hence, this suggests that the transcribing quality is significantly high. We consider the annotations as it is for the analysis since the transcriptions include the accented phonemes that could indicate phoneme sound quality besides phoneme mispronunciation.

4. DATA ANALYSIS

4.1. Setup

We analyse the effectiveness of the annotated transcriptions by exploring the dependencies between the errors made by Indian L2 learners with the overall ratings and factor specific binary scores. The errors are computed by considering canonical pronunciations of UK accented English as the ideal pronunciation reference, for which we use BEEP dictionary [18].

We consider a set of five measures representing the errors for the analysis. To compute these measures, we consider alignment [19] between Indian L2 speaker's phoneme transcription and UK English canonical transcription for each utterance and obtain phoneme errors (includes phoneme insertions, deletions and substitutions) in the utterance respect to canonical transcription. The first measure is the phoneme errors in an utterance averaged across all utterances, referred to as PE. The second measure is the phoneme errors per phoneme (where the denominator is the total number of phonemes in the canonical pronunciation) in an utterance averaged across all utterances. This measure is popularly known as phoneme error rate (PER) in the ASR literature. The third measure is the phoneme errors per word in an utterance averaged across all the utterances, referred to as PE per word. We believe that the PE measure is an absolute error in an utterance. Thus, it depends on the number of phonemes in the utterance; hence we consider PER and PE per word. The PE per word is considered based on the hypothesis that single or multiple phoneme mispronunciations in a word can affect the rating similarly. Further, this hypothesis is considered the basis for choosing the fourth and fifth measures. The fourth measure is the number of phoneme-mispronounced words (we consider any single or multiple phoneme mispronunciations in a word as a phoneme-mispronounced word) in an utterance averaged across all the utterances, referred to as PMW. The fifth measure is PMW per word; typically, this measure resembles the rate (WER) in the ASR literature and is referred to as phoneme mispronounced word rate (PMWR).

Considering the five measures, we analyse the effectiveness of the annotations in the following three aspects: 1) exploring the dependencies between all the five measures with overall rating, 2) exploring the dependencies between all the five measures with the factor specific binary scores, and 3) computing correlation between the each of the five measures with the overall rating and all the seven factor specific binary scores. All three aspects are analysed separately considering the following four data conditions: 1) entire data, 2) data containing single words, 3) data with short sentences, and 4) data with long sentences. We exploit the dependencies considering an average (first-order statistics) trend in the respective data conditions separately from the first two aspects. On the other hand, the third analysis aspect exploits the second-order statistics trend in the respective data separately.

4.2. Outcomes

4.2.1. With overall rating:

Figure 1 shows the values of PE, PER, PE per word, PMW and PMWR for each overall rating under all four data conditions. The figure shows that the PE is decreasing with overall ratings from 1 to 10 except for rating 0 when we consider the entire data. However, the PER and PE per word have a decreasing trend from rating 0 to 10. This could be because the absolute error is susceptible to the number of phonemes or words in the considered utterances when the entire data consists of variable length sentences. While observing the trend separately for single-word, short and long sentences, the gap between PE at rating 0 and rating 1 is low compared to that observed in the entire data condition. This supports the susceptible nature of absolute errors to the sentence length. Further, the PE, PER and PE per word decrease from low to high ratings in most cases separately for single-word, short and long sentences. The observations based on PE, PER and PE per word follow similar trends while comparing PMW and PMWR under all four data conditions. The decreasing trend in all the cases in the figure supports the use of annotated phoneme transcriptions for building the pronunciation assessment and the mispronunciation detection and diagnosis models.

[1] Link to details on IPA Symbols used

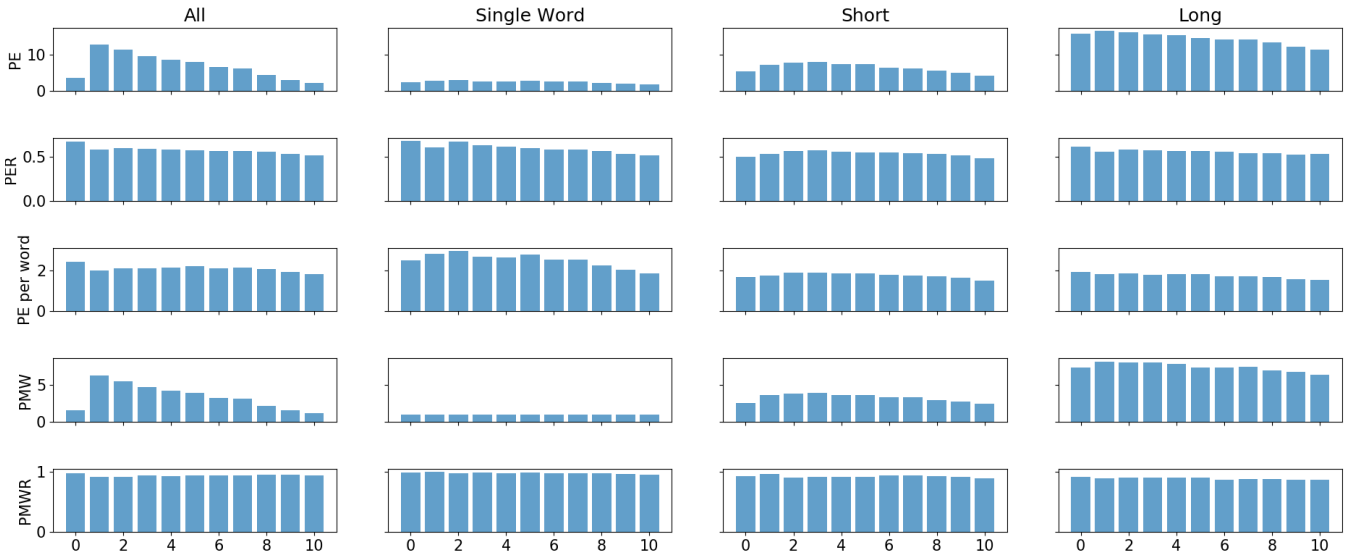

Fig. 1. The values of PE, PER, PE per word, PMW and PMWR for each overall rating under the four data conditions.

4.2.2. With factor specific binary scores:

Figure 2 shows the values of PE, PER, PE per word, PMW and PMWR for class 0 and 1 of each factor under all four data conditions. We present the analysis by grouping the factors for better readability.

Absence of mispronunciation: Among all the seven factors, it is expected that the PE and PMW inversely relates with the absence of phoneme mispronunciation. The figure shows that the average values of all the five measures are high for class 0. i.e. presence of mispronunciation and vice versa under all four data conditions. This indicates that the annotated phonetic transcriptions are rich in quality. Also, this suggests the usage of phonetic transcriptions for mispronunciation detection and diagnosis applications. While observing the remaining six factors, the relation between class 0 and class 1 is the same across all the measures separately for each factor under all four data conditions.

Phoneme quality, Stress, and Chunking: The values of the measures are high for class 0 and low for class 1 for all these three factors under all the four data conditions. The higher values of the measures for class 0 under the phoneme quality could be because the phonetic transcriptions also include the symbol specific to the accent. Thus, the accented phoneme symbols also contribute to the errors; hence the inverse relationship exists. Further, under Stress, the higher values for class 0 could be because the phoneme errors might include the substitutions of short vowels with long vowels and vice versa, as well as diphthongs with long/short vowels, mainly in the Indian context. Thus, these replacements result in more PE or PMW and reduce the stress quality. Further, the increased PE or PMW could indicate a higher cognition load, which would cause improper chunking. Hence, annotated phoneme transcriptions can be used for more detailed investigations on the relationship between these three factors and errors.

Intonation: The values of the measures are almost similar between class 0 and 1 under all the conditions. This is because the intonation mainly depends on the tonal variations, which may not be affected due to the errors.

Intelligibility and Absence of MTI: The values of the measures are low for class 0 and high for class 1 for the two factors under all the four data conditions. This is because both the factors depend on the cumulative phenomena of all the remaining factors and the errors. Hence, direct relations may not be drawn between these factors and all the considered measures. However, we believe that more detailed investigations are required on the conclusion about the relationship between both the factors and the errors. The annotated transcriptions in this work would allow the researchers to explore the above directions.

4.2.3. Correlation with overall rating and factors' scores:

Figure 3 shows the correlation obtained between the values of each of PE, PER, PE per word, PMW and PMWR with the overall rating and factors' scores under the four data conditions. From the figure, among all the seven factors and overall rating, the overall ratings highly correlated (negatively) with all the five measures except PMWR under all four data conditions. This further supports the previous claim that the annotated phoneme transcriptions are of rich quality for the task of pronunciation assessment. Among the seven factors, the absence of mispronunciation (#3) highly (negatively) correlated with all the five measures except PMWR under all four data conditions. Comparing the remaining six factors, phoneme quality (#2), stress (#4) and chunking (#6) are negatively correlated with all the five measures in most cases under all the

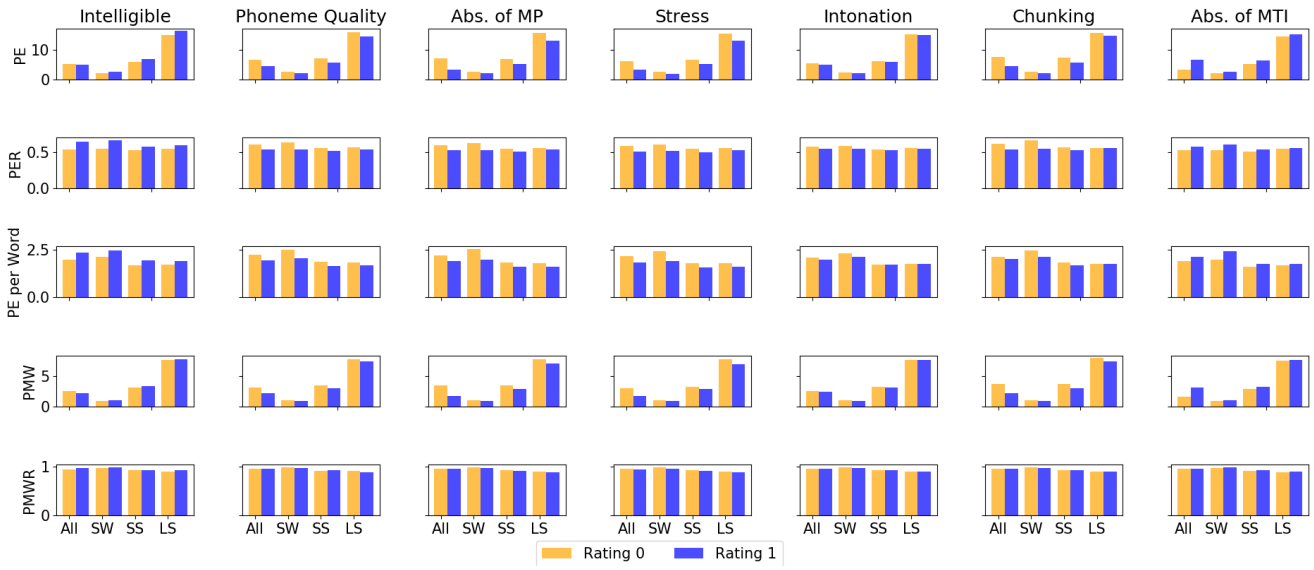

Fig. 2. Histogram of PE, PER, PE per word, PMW and PMWR for class 0 and 1 for all the seven factors under the four data conditions. (All- Entire data, SW- Single Words, SS- Short Sentences and LS- Long Sentences)

four data conditions. The correlation obtained for intonation quality (#5) is the lowest for all the five measures among all the seven factors, and it is close to zero for most of the measures. Hence, there is no correlation between intonation and the errors. Finally, intelligibility (#1) and absence of MTI (#7) have a positive correlation when all the five measures under all the four data conditions of the data. These observations are similar to those based on Figure 2.

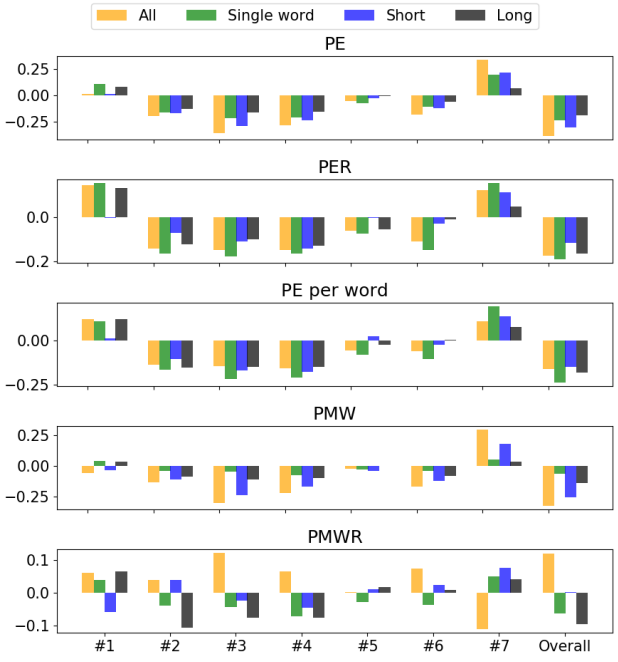

Fig. 3. Spearman correlation coefficient of PE, PER, PE per word, PMW and PMWR with the factor specific scores and overall rating under the four data conditions.

5. CONCLUSION

We develop voisTUTOR 2.0 by including phonetic annotations to the audios in the existing voisTUTOR 1.0, considering a linguist with approximately 15 years of experience. This makes voisTUTOR 2.0 a unique corpus consisting of 26259 utterances from Indian L2 learners counting approximately 14 hours with annotations of phones, pronunciation quality ratings and binary scores of the factors influencing the pronunciation quality for each audio. We believe that voisTUTOR 2.0 could be useful for building robust models for CAPT applications include pronunciation assessment and mispronunciation detection and diagnosis by exploring interdependencies between pronunciation errors and ratings. A preliminary analysis on voisTUTOR 2.0 suggests that a set of five measures indicative of the errors correlated with the pronunciation quality ratings and the binary scores indicating mispronunciations and phoneme quality. Further developments are required to obtain ratings, binary decisions and phonetic annotations from multiple experts and to evaluate consistency across the raters and the annotators separately.

6. ACKNOWLEDGEMENT

We would like to acknowledge Dr. Sharmistha Charakrabarti, a Linguist at Spire Lab, Electrical Engineering, Indian Institute of Science, for transcribing all the audio files in our Data and in turn making this analysis possible.

7. REFERENCES

[1] Gwo-Jen Hwang and Qing-Ke Fu, "Trends in the research design and application of mobile language learning: A review of 2007–2016 publications in selected ssci journals," *Interactive Learning Environments*, vol. 27, no. 4, pp. 567–581, 2019.

[2] Nina Hosseini-Kivanani, Roberto Gretter, Marco Matassoni, and Giuseppe Daniele Falavigna, "Experiments of asr-based mispronunciation detection for children and adult english learners," *arXiv preprint arXiv:2104.05980*, 2021.

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

[4] Wei Li, Sabato Marco Siniscalchi, Nancy F Chen, and Chin-Hui Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 6135–6139.

[5] Savitha Murthy, Ankit Anand, Avinash Kumar, Ajay Cholin, Ankita Shetty, Aditya Bhat, Akshay Venkatesh, Lingaraj Kothiwale, Dinkar Sitaram, and Viraj Kumar, "Pronunciation training on isolated kannada words using" kannada kali"-a cloud based smart phone application," in *2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*. IEEE, 2018, pp. 57–64.

[6] Joyanta Basu, Soma Khan, Rajib Roy, Babita Saxena, Dipankar Ganguly, Sunita Arora, Karunesh Kumar Arora, Shweta Bansal, and Shyam Sunder Agrawal, "Indian languages corpus for speech recognition," in *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2019, pp. 1–6.

[7] Chiranjeevi Yarra, Aparna Srinivasan, Chandana Srinivasa, Ritu Aggarwal, and Prasanta Kumar Ghosh, "voistutor corpus: A speech corpus of indian l2 english learners for pronunciation assessment," in *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2019, pp. 1–6.

[8] Horacio Franco, Harry Bratt, Romain Rossier, Venkata Rao Gadde, Elizabeth Shriberg, Victor Abrash, and Kristin Precoda, "Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

[9] Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang, "Speechocean762: an open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[10] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub, "Automatic scoring of pronunciation quality," *Speech communication*, vol. 30, no. 2-3, pp. 83–93, 2000.

[11] Ananlada Chotimongkol, Sumonmas Thatphithakkul, Patcharika Chootrakool, Chatchawarn Hansakunbuntheung, and Chai Wutiwiwatchai, "The design and development of pelecan: Pronunciation errors from learners of english corpus and annotation," in *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*. IEEE, 2011, pp. 36–41.

[12] Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus.," in *INTERSPEECH*, 2018, pp. 2783–2787.

[13] Yu Chen, Jun Hu, and Xinyu Zhang, "Sell-corpus: an open source multiple accented chinese-english speech corpus for l2 english learning assessment," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7425–7429.

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

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

[16] Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter, "The isle corpus of non-native spoken english," in *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*. European Language Resources Association, 2000, pp. 957–964.

[17] Jacob Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[18] T Robinson, "Beep dictionary," *BEEP dictionary*, 1996.

[19] Gonzalo Navarro, "A guided tour to approximate string matching," *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.