A temporal analysis of spoken language learning for automatic tutoring

Chiranjeevi Yarra

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

{chiranjeeviy}@iisc.ac.in

1. Objective

Spoken language (SPL) learning especially English language in India has a lot of demand in achieving language proficiency. SPL training with an automated language learning tool is cost effective compared to an human based training. In this work, we attempt to build such an automated tool for Indian learners learning English. The proposed tool would train the learners by evaluating their performance and providing feedback to them in an interactive manner like a human expert.

2. Summary

2.1. Motivation

Spoken language learners, specially English language learners, are often influenced by their nativity. For example, the prosodic rules in the English language pronunciation are often distorted by the native prosodic rules [1]. These nativity influences are required to be minimized for achieving a good quality pronunciation. Effective spoken language training methods can compensate such nativity influences. However, these methods are expensive as it requires highly proficient English experts [2]. In cases, where the cost effective solutions are required, it is useful to have an automated system which performs tasks of detecting the mispronunciation and providing feedback to minimize nativity influences like an human expert. Such kind of systems are also useful for people who can not avail the high quality training methods due to their demographic and physical constraints. Kam et al. have studied that the rural Indian people, which comprise the majority of Indian population, can not avail the high quality educational resources but they are improving their learning via information communication channels (ICC) like mobiles, Internet and computer interfaces [3]. However, it is still a challenge to develop such an automated system for providing feedback in an interactive manner using ICCs.

Most of the existing works have addressed the problem of automatic identification of pronunciation quality [1, 2, 4–7]. However, in these works, the chosen learners do not come from Indian nativity. There are few works that have addressed the problem of providing [8, 9]. However, those are restricted to a small set of errors, e.g., phonemic errors among many types of pronunciation errors. On the other hand, in Indian nativity, few existing works have addressed either methods for automatic pronunciation assessment [10] or phonological studies for the benefit of pronunciation assessment [11]. However, none of the existing works have addressed the problem of automatic feedback for improving pronunciation quality in Indian nativity.

2.2. Proposed methodology

India has a large diversity in the nativities. Phonetic rules varies largely across these nativities. Thus, Indian English many accents across the India. Analyzing all the accent variabilities and developing an automatic feedback system is tedious. In order to circumvent this problem, we propose to train the learners with the received pronunciation or neutral accent of English. Also, irrespective learners nativity, we estimate the learners' performance in reference to one male and one female expert as in a class-room teaching set-up i.e., one teacher versus many students. Further, it has been observed that the estimation of a learner's performance would be more accurate when automatic speech recognition (ASR) engine built with Indian accent speech data [10]. Hence, in this work, we propose to build ASR models using the speech data collected from Indian subjects belonging to different native languages.

Typically, the pronunciation quality depends on the degree of the following sub-qualities – 1) phoneme quality 2) lexical stress quality 3) intonation quality 4) fluency [12]. Thus, we consider to train the learners at these four levels and propose an automated system. The proposed system displays a set of stimuli sequentially at each level and records the audio of the stimuli from learner. Afterwards, it analyzes the recorded audio and a performance score is displayed along with audio visual feedback. Based on these the learners can decide to choose appropriate practice aids to improve their pronunciation skill. The proposed system automatically generates the feedback specific to each sub-quality along with a performance score. Hence, it is required to build a database and validate the proposed system at each sub-quality level. However, the existing corpora [4, 7, 13–16] do not contain rich set of utterances from Indian learners with expert ratings of each sub-quality.

2.3. Work done

2.3.1. Data collection

We collect two sets of data one for building ASR models, referred to as ASR data, and another for assessment and feedback models, referred to as pronunciation data. For the ASR data, we record 2304 TIMIT sentences from 80 subjects from different native places within India with varying proficiency in English pronunciation. This comes to a total of 200 (approx) hours of data. For the pronunciation data, we consider stimuli specific to each sub-quality as follows – 1) for the phoneme quality, a total of 1023 word stimuli containing minimal pairs and small to large number of consonant clusters, 2) for the stress quality, a total of 230 stimuli containing words and small phrases, 3) for the intonation quality, a total of 80 stimuli containing phrases and 4) for the fluency, a total of 343 stimuli containing simple to complex sentences. A total of 1676 stimuli are recorded from 16 subjects from different native places within India. At the time recording, all the 16 subjects were engaged in spoken English training classes. We also record these stimuli from two experts (one male and one female), who are voice-over artists besides that the female speaker is a spoken English teacher.

2.3.2. Modeling

Phoneme quality: It is effected by phonetic errors in the pronunciation such as insertions, deletions and substitutions and the degree of accent in the phonetic sounds. In most of the cases, these errors and the accented phonetic sounds cause loss of intelligibility in the pronunciation. In order to provide feedback on these errors, we propose to identify the errors by performing a force-alignment on the recorded audio using a pronunciation lexicon consists of Indian specific pronunciation errors. Where, the lexicon is proposed to construct using the rule set following the work by Sailaja [17]. Further, in order to reduce the accent in phonetic sounds, we propose to show the videos consists of articulatory movements made by the expert for the stimuli specific to this sub-quality. However, creation of such videos involves specialized equipment which are costly and time consuming. We also observed that the equipments involve physical contact with the articulators cause degradation of the audio quality [18]. In order to circumvent these, we consider

real-time magnetic resonance imaging (rt-MRI) videos [19] and synthesize expert videos for all 1023 stimuli using image frames from rt-MRI videos and the respective expert's audio [20, 21].

Lexical stress quality: A good quality pronunciation is achieved with correct placement of the stress markings on the syllables. We hypothesize that detecting the stress markings made by learners and provide those along with stress markings made by the experts as a feedback could benefit the learners. Typically, the stress detection task is posed as a classification problem in a supervised manner, where a set of features are derived for each syllable and used along with manually marked labels – stressed and unstressed. In the feature computation, we propose to incorporate sonority in short-time energy and show the absolute improvement of 3.09% in stress detection accuracy [22] compared to the baseline [23] considering only short-time energy. It has been shown that the features are not only vary with stress markings but also with context [23, 24]. In order to normalize the context based effects, we perform stress detection in two ways. First, we add context as binary features to the sonority based features and allow classifier to learn the normalization properties implicitly. Second, we learn a decision tree using context and predict stress markings at all the nodes in the tree that are active during test time. Considering all these markings, we propose to obtain a final stress marking for each syllable. Between the two methods, the second method achieves the highest accuracy of 92.17%, which is an absolute improvement of 8.92% compared to the baseline [23].

The quality of the stress detection task depends on the labels and the syllable data (boundaries and transcriptions), which are typically obtained using manual annotation. However, in the proposed system during testing, syllable data is estimated using force-alignment process which in turn could cause mismatch between estimated and annotated data. Thus the classifier train with annotated data could degrade the stress detection performance under test condition. In order to circumvent this, we propose to map labels of the annotated data onto the estimated data instead of obtaining new labels for the estimated data manually. In general, the manual labeling and the annotation are time consuming and cost ineffective. Thus, we propose to formulate the stress detection task in an unsupervised manner.

Intonation quality: Typically, spoken English learners are trained to produce four different types of intonation referred to as intonation classes. These classes can be distinguished using temporal structures in the pitch in an utterance. Thus, the accuracy in the pitch estimation reflect the quality of intonation. In order to estimate pitch more accurately compared to existing techniques, we propose a dynamic programming based pitch estimation to ensure smooth pitch transitions within the voiced segments [25]. Following this, we propose a model to detect intonation class considering temporal structures in the estimated pitch in a supervise manner using hidden Markov model (HMM) and long short term memory (LSTM) networks. Further, in order to reduce cost involved in manual labeling, we propose an unsupervised approach for intonation detection. The proposed supervised and unsupervised methods achieves improvement (absolute) of 30% and 39% compared to the baseline [26]. We hypothesize that the learners could be benefited by providing detected intonation of their utterance as a feedback. Further, they could be more benefited by showing temporal structures in the pitch with a sequence line segments, which is a typical training practice in intonation learning [27]. We obtain these by stylizing the estimated pitch considering absolute sum as the cost function. Further, we show that the absolute sum based stylization is robust to pitch estimation errors compared to traditional root mean square error based stylization [28].

Fluency: Most of the works have shown that the fluency can be measured using speaking rate [29]. Among many measures, it has been shown that the syllable rate (number of syllables per second) is mostly correlated with fluency [30]. For this, we propose methods to compute syllable rate directly from speech acoustics [31, 32] instead of estimating syllables using ASR models. The proposed method achieves improvement in estimating syllable rate in both clean and noisy conditions compared to the baseline [30]. Further, based on the syllable rate, we propose a set of measures to know its effectiveness along with acoustic and prosodic based measures in identifying nativity of the learners. In addition to the speaking rate, the learners are required to adapt themselves to speech rhythm in English, where the rhythm is the extent to which there is a regular beat in speech. We observe that the languages in India are rhythmically well discriminated from English. Hence, learners speech with more influence of native rhythm causes poor pronunciation quality. Thus, providing speaking rate and degree of native rhythm influence in learners utterance as a feedback could help them to achieve fluency. Further, the learners could obtain better fluency by correctly chunking the utterance. This can be trained by showing pauses made by the learners in respect to correct pauses made by the experts.

Performance score estimation: Typically, in the literature, performance of the learners has been estimated at discrete levels, for which, classification based approaches have been considered in a supervised manner [7, 12, 33]. Similar to the existing works, we, in this work, propose to estimate Indian learners performance at ten discrete levels. For this, we obtain markings of these levels for the pronunciation data from an expert. Further, in order evaluate the quality of the feedback at all four sub-qualities, we propose to obtain binary ratings (1 for Yes and 0 for No) for the following six yes or no queries – 1) any loss of intelligibility 2) is phoneme sound quality is good 3) any incorrect stress markings 4) is intonation is correct 5) any incorrect placement of pauses 6) any mother tongue influence. Though the queries do not match with the proposed feedback in the four sub-qualities, we hypothesize that one or more combinations of these queries would reflect the proposed feedback indirectly. Further, we believe that the score prediction model could be more effective when it estimates the binary ratings in addition to the learners performance.

2.3.3. Interface design

We, in this work, design a tutoring system in a web based interface, where front-end is available at learners location and back-end is situated in server at our end. At the front-end, a stimuli to be read is displayed and it records learners utterance. Once the recording is over, the entire audio is streamed to server in packets [34]. This is done to avoid distortion of the audio as in typical real-time communication scenarios, where the data undergoes encoding and decoding process [35]. We have shown that the encoding and decoding of data causes quality degradation in speech applications [35]. So far, in the proposed web interface, we incorporate modules of the proposed feedback of stress and intonation quality [36, 37].

2.4. Further work

The research plan for the thesis is as follows:

- Estimation of syllable stress in an unsupervised way.
- Estimation of the performance score along with the binary ratings.
- Incorporation of the proposed feedback for all qualities.

3. References

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

[2] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech & Language*, vol. 25, no. 2, pp. 282–306, 2011.

[3] M. Kam, A. Kumar, S. Jain, A. Mathur, and J. Canny, "Improving literacy in rural india: Cellphone games in an after-school program," *International Conference on Information and Communication Technologies and Development (ICTD)*, pp. 139–149, 2009.

[4] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning," *Proceedings of InSTILL*, pp. 123–128, 2000.

[5] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," *International Conference on Spoken Language*, vol. 3, pp. 1457–1460, 1996.

[6] J. Tepperman and S. Narayanan, "Hidden-articulator markov models for pronunciation evaluation," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 174–179, 2005.

[7] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody - annotation, modelling and evaluation," *International Symposium on Automatic Detection of Errors in Pronunciation Training*, pp. 21–30, 2012.

[8] K. Li and H. Meng, "Mispronunciation detection and diagnosis in L2 english speech using multi-distribution deep neural networks," *International Symposium on Chinese Spoken Language Processing*, pp. 255–259, 2014.

[9] 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.

[10] S. Joshi and P. Rao, "Acoustic models for pronunciation assessment of vowels of indian english," pp. 1–6, 2013.

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

[12] V. Ramanarayanan, P. Lange, K. Evanini, H. Molloy, and D. Suendermann-Oeft, "Human and automated scoring of fluency, pronunciation and intonation during human–machine spoken dialog interactions," *Interspeech*, pp. 1711–1715, 2017.

[13] E. Atwell, P. Howarth, and D. Souter, "The ISLE corpus: Italian and german spoken learner's english," *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, vol. 27, pp. 5–18, 2003.

[14] E. Grabe, B. Post, and F. Nolan, "The IViE corpus," *Department of Linguistics, University of Cambridge*, 2001.

[15] J. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos, "The HIWIRE database, a noisy and non-native english speech corpus for cockpit communication," *Online. http://www. hiwire. org*, 2007.

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

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

[18] N. Meenakshi, C. Yarra, B. Yamini, and P. K. Ghosh, "Comparison of speech quality with and without sensors in electromagnetic articulograph ag 501 recording," *Interspeech*, pp. 935–939, 2014.

[19] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time MRI articulatory corpus for speech research," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[20] U. Desai, C. Yarra, and P. K. Ghosh, "Concatenative articulatory video synthesis using real-time MRI data for spoken language training," *Accepted in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. –, 2018.

[21] S. Chandana, C. Yarra, R. Aggarwal, S. K. Mittal, N. K. Kausthubha, K. T. Raseena, A. Singh, and P. K. Ghosh, "Automatic visual augmentation for concatenation based synthesized articulatory videos from real-time MRI data for spoken language training," *Accepted in Interspeech2018*, pp. –, 2018.

[22] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5845–5849, 2017.

[23] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners." *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 937–940, 2005.

[24] R. Delmonte, M. Petrea, and C. Bacalu, "Slim prosodic module for learning activities in a foreign language," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[25] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "A frame selective dynamic programming approach for noise robust pitch estimation," *The Journal of the Acoustical Society of America*, vol. 143, no. 4, pp. 2289–2300, 2018.

[26] K. Li, X. Wu, and H. Meng, "Intonation classification for L2 English speech using multi-distribution deep neural networks," *Computer Speech & Language*, vol. 43, pp. 18–33, 2016.

[27] J. D. O'Connor, "Better english pronunciation," *Cambridge University Press*, 1980.

[28] P. K. Ghosh and S. S. Narayanan, "Pitch contour stylization using an optimal piecewise polynomial approximation," *IEEE signal processing letters*, vol. 16, no. 9, pp. 810–813, 2009.

[29] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[30] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[31] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection," *Speech Communication*, vol. 78, pp. 62–71, 2016.

[32] S. Nagesh, C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "A robust speech rate estimation based on the activation profile from the selected acoustic unit dictionary," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5400–5404, 2016.

[33] J. Tepperman, J. Silva, A. Sethy, and S. Narayanan, "Robust recognition and assessment of nonnative speech variability," *International Conference on Intelligible Systems and Computations: Theory and Applications*, 2006.

[34] K. R. Fall and W. R. Stevens, *TCP/IP illustrated, volume 1: The protocols*. addison-Wesley, 2011.

[35] S. Raghavan, N. Meenakshi, S. K. Mittal, C. Yarra, A. Mandal, K. P. Kumar, and P. K. Ghosh, "A comparative study on the effect of different codecs on speech recognition accuracy using various acoustic modeling techniques," *National Conference on Communications*, pp. –, 2017.

[36] C. Yarra, P. A. Anand, N. K. Kausthubha, and P. K. Ghosh, "SPIRE-SST: An automatic web-based self-learning tool for syllable stress tutoring (SST) to the second language learners," *Accepted in Interspeech2018*, pp. –, 2018.

[37] P. A. Anand, C. Yarra, N. K. Kausthubha, and P. K. Ghosh, "Intonation tutor by SPIRE (In-SPIRE): An online tool for an automatic feedback to the second language learners in learning intonation," *Accepted in Interspeech2018*, pp. –, 2018.