INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

SPIRE-fluent: A self-learning app for tutoring oral fluency to second language English learners

Chiranjeevi Yarra[1], Aparna Srinivasan[1], Sravani Gottimukkala[2], Prasanta Kumar Ghosh[1]

[1]Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India
[2]Rajiv Gandhi University of Knowledge Technologies (RGUKT), Kadapa, 516330, India
{chiranjeeviy,aparnasrinivasa}@iisc.ac.in, sravaniigottimukkala@gmail.com, prasantg@iisc.ac.in

Abstract

Second language (L2) learners often achieve oral fluency by correct pronunciation of words with appropriate pauses. It has been shown that the L2 learners improve their language skills using mobile apps in a self-learning manner. Effective learning is possible with apps that provide detailed feedback. However, apps that train oral fluency in an automatic way are not available. In this work, we present SPIRE-fluent app, which provides an automatic feedback with scores representing learner's pronunciation quality, for each word in a sentence and for the entire sentence. The word specific scores are computed based on the correctness of pronunciation with respect to the expert's audio. Further, the app displays the syllables uttered and a set of two types of pauses produced by the learners and the expert while speaking the sentence. Considering this as a feedback, the learner can correct their mistakes based on the mismatches between those utterances. In addition, it also estimates any pause made by the learners within a word and highlights the syllable containing the phoneme preceding the pause.

1. Introduction

Generally, the oral fluency of L2 learners is considered as a measure of language proficiency, and it can be improved by incorporating proper pause placement and correct pronunciation of words. With the advent of computer-aided language learning (CALL), mobile apps that provide automatic feedback on oral fluency can act as effective self-learning tools for L2 learners [1]. Even though there exist multiple associated works, apps that provide such feedback to the learner are limited. For the benefit of L2 learners we have developed SPIRE-fluent app for improving the sentence speaking skills by providing an automatic feedback on pause locations and the pronunciation quality of each word in a sentence uttered by the learner. The pause locations and the pronunciation quality are estimated automatically in real time. Moreover, it has been shown that such online tools benefit the learners, who otherwise do not have easy access to similar training methods [1]. To the best of our knowledge, no similar app exits.

2. Proposed Architecture

Figure 1 shows the proposed architectures of SPIRE-fluent app, which comprises of a front end and a back end. The front end consists of user interfaces (UIs) that are accessed with a mobile, through which the learner can interact. On the other hand, the recorded audio is processed and scored in the back-end server at our side. The front end and the back end communicate with one another via the internet.

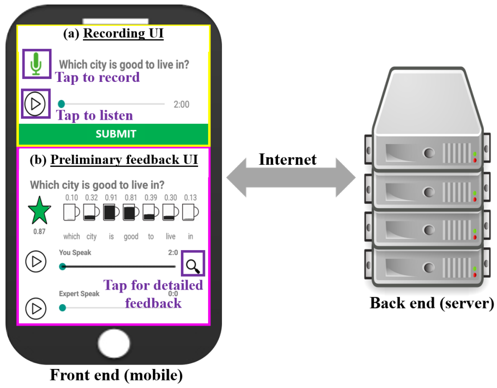

Figure 1: *The architecture of the SPIRE-fluent app.*

2.1. Front end

The front-end of SPIRE-fluent has three UIs and each UI is dedicated to one of the three main functionalities namely, 1) recording, 2) preliminary feedback and 3) detailed feedback. Each of these functionalities are explained in detail in the following subsections.

2.1.1. Recording UI

On launching the app, the learner can choose one among four lessons, where each lesson contains multiple stimuli for practice. The screenshot a in Figure 1 shows an exemplary UI that appears once a lesson is selected. In the UI, a stimuli to read is displayed, and beside the stimuli a microphone symbols is provided. On tap of the microphone symbol, the learner can start recording his/her audio and simultaneously the symbol changes to a stop button, which can be tapped to terminate the recording. After completion of recording a play button appears below the stimuli, which can be used to listen to the latest recording. Thus, the learner can continue recording until satisfied and send the latest recording to the back end by tapping the "SUBMIT" button.

2.1.2. Preliminary feedback UI

On tapping the "SUBMIT" button the word level and the sentence level scores appear as shown in screenshot b in Figure 1. In the UI, each word in the sentence is displayed below a mug. Partially filled mugs are used to represent the word level scores indicating the pronunciation quality of a learner, and the scores are displayed above the mugs. Further, in the UI there is a provision to listen to the expert's audio and learner's recorded audio by tapping the respective play buttons. By listening to the expert's audio, the learner can correct their pronunciation mistakes. However, if the learner fails to do so, he/she can view the detailed feedback by tapping the magnifying glass symbol near the learner's audio recording, with which we believe that the learner can correct their mistakes easily.

Detailed feedback UI

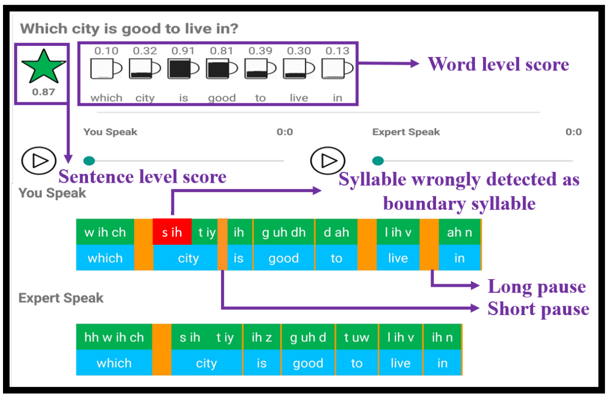

Figure 2: *A screenshot of an exemplary detailed feedback UI.*

2.1.3. *Detailed feedback UI*

Figure 2 shows an exemplary UI displaying the detailed feedback. In this UI, in addition to the pronunciation quality scores, it displays syllables and pauses present in the learner's and expert's utterances for the given stimuli. A set of two pauses are considered as indicative of two types of boundaries – 1) intermediate phrase boundary and 2) intonation phrase boundary [2], and are referred to as short pause and long pause respectively. The syllable preceding a pause is known as a boundary syllable. In the UI, the location of the pauses are represented by orange strips, where broad and thin strips denote the long and short pauses respectively. Further, the UI highlights the syllables uttered by the learners with red and green colors. A syllable is highlighted in red, if it is not a word final syllable, but it gets detected as a boundary syllable. Otherwise the syllable is highlighted in green.

2.2. Back end

Given an audio, in order to estimate the syllables, we perform forced-alignment to estimate the phoneme transcriptions and its boundaries. These transcriptions are syllabified to obtain syllable transcriptions from which, the syllable boundaries are determined. Following this, we identify the syllables contained in each word. In addition, to detect pauses, we compute features for each syllable and classify it as one of the three classes – 1) boundary syllable associated with short pause, 2) boundary syllable associated with long pause and 3) not a boundary syllable. Further, the probability associated with each class is considered for score computation.

2.2.1. *Score computation*

The word level score is obtained by considering the goodness of pronunciation (GoP) measure [3] for each phoneme p within the word, normalized using complete phoneme set Q. Considering these, the score is computed as, $1 - \tanh(\alpha|n^E - n^L|)$, where E and L represent expert and learner, n is the average of $\frac{\text{GoP}(p)}{\sum_{q \in Q} \text{GoP}(q)}$ across all the phonemes in the word and α is set to 2. Following this, the sentence level score is obtained by averaging the word level scores and the scores representing correctness of the pauses associated with the syllables in each word. For each word, the pause based score is considered as zero when any syllable other than word final syllable is detected as boundary syllable. Otherwise, the pause based score for every word is the probability of the word final syllable belonging to the class, to which the corresponding syllable in the expert's utterance is assigned.

3. Demonstration

For the demo, we perform forced-alignment using Kaldi speech recognition tool kit [4] and a lexicon containing pronunciations for each word and. We consider P2TK syllabifier [5] for syllabification. We estimate the two types of pauses in both expert's and learner's utterances following the work by Ananthakrishnan et al. [2] and it is implemented along with score computation using Python programming. We use Android SDK for developing the front end. We set-up the server using LAMP (Linux, Apache, MySQL, PHP) stack on Ubuntu 14.04 LTS operating system. The stimuli available in the SPIRE-fluent are taken from the material used for spoken English training [6]. We obtain the expert's audio by recording the stimuli from a voice-over artist, proficient in British English spoken communication.

4. Conclusion

We present the Android app SPIRE-fluent, that teaches L2 English learners the nuances of oral fluency. We design the front end with Android SDK and back end codes with Python programming language. Using this app, we provide feedback by providing word and sentence level scores and by showing two types of pauses and the syllables produced by the learner and expert in their pronunciations. Further investigations are required to measure the effectiveness of the proposed app as well as analyze sufficiency of the feedback parameters in the self-learning process.

5. Acknowledgement

We thank the Department of Science & Technology, Government of India and the Pratiksha Trust for their support.

6. References

[1] A. Kumar, A. Tewari, G. Shroff, D. Chittamuru, M. Kam, and J. Canny, "An exploratory study of unsupervised mobile learning in rural India," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 743–752.

[2] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 216–228, 2008.

[3] W. Hu, Y. Qian, and F. K. Soong, "An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech." in *SLaTE*, 2015, pp. 71–76.

[4] D. Povey, *et al.*, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[5] J. Tauberer, "P2TK automated syllabifier," *Available at https://sourceforge.net/p/p2tk/code/HEAD/tree/python/syllabify/*, last accessed on 24-04-2019.

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