voisTUTOR: Virtual Operator for Indian Spoken English TUTORing

Chiranjeevi Yarra, Prasanta Kumar Ghosh

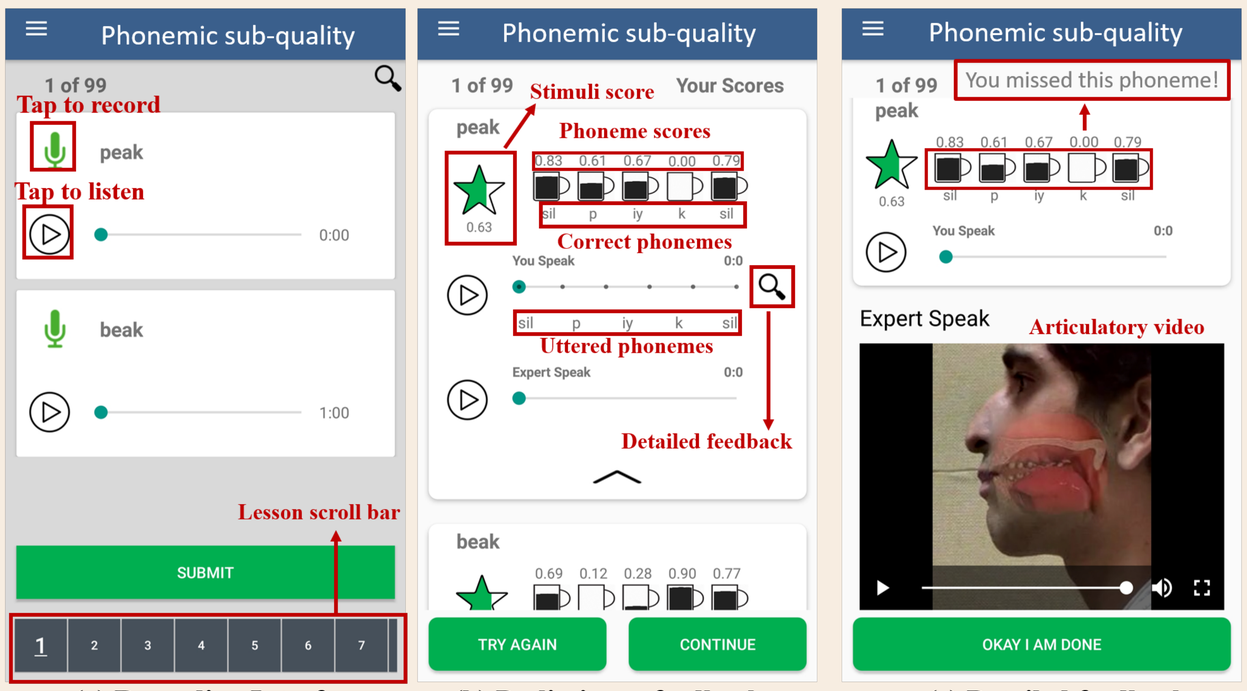
Electrical Engineering, Indian Institute of Science (IISc), Bangalore.

PROBLEM STATEMENT

Demonstrate voisTUTOR tool that provide detailed feedback on correct usage of phonemes as well as stress, intonation and pauses for Indian learners to learn spoken English pronunciation in a self-learning manner.

Phoneme

- **Why to learn usage of phonemes:** Correct pronunciation of phonemes is essential for second language (L2) learners to be able to speak words/sentences like a native speaker [1].



(a) Recording Interface (b) Preliminary feedback (c) Detailed feedback

- **Score computation:** Compute the scores for following phoneme categories: 1) insertions $\left(S^i_{avg}\right)$, 2) deletion $\left(S^d_{avg}\right)$, 3) substitution $\left(S^s_{avg}\right)$ and 4) remaining $\left(S^r_{avg}\right)$.
- These phoneme categories are estimated with forced-alignment.
- The overall score for a word (w) is computed as $S(w) = \frac{\sum_{j\in\{i,d,s,r\}} \alpha_j S^j_{avg}}{\sum_{j\in\{i,d,s,r\}} I_j \alpha_j}$, where $I_j \in \{1, 0\}$. $I_j = 1$ when j-th category occurs in the word (w) and α_j is a weight associated with j-th category.

Syllable stress

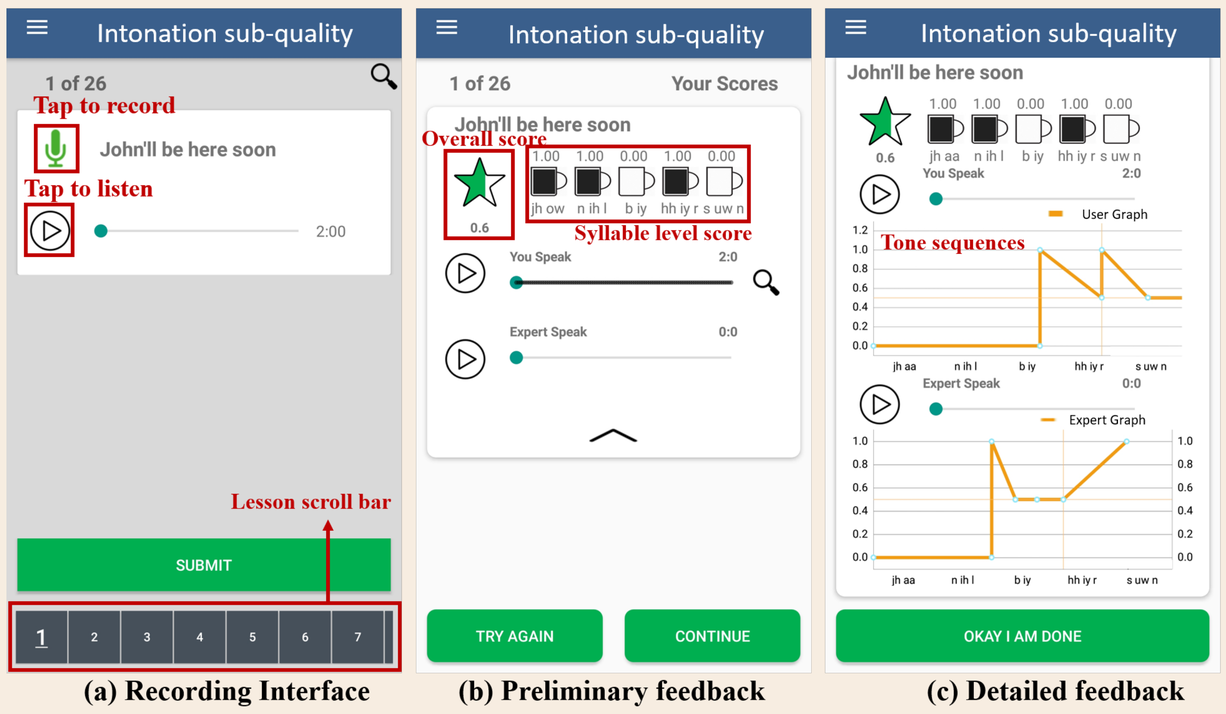
- **Why to learn usage of syllable stress:** In the language learning, correct usage of syllable stress patterns could minimize the localized pronunciation errors [1].

(a) Recording Interface (b) Preliminary feedback (c) Detailed feedback

- **Syllable stress detection:** Using automatic speech recognition toolkit and syllabification software, estimate syllable transcriptions and its time-aligned boundaries.
- Following work by Yarra et al.[2], for each syllable, estimate the stress markings as well as scores representing the confidence in estimating those markings.
- Using the stress markings and the confidence scores, pronunciation quality score is computed.
- **Score computation:** $S_E(i)$ and $S_L(i)$ are the scores corresponding to the expert and the learner for i^{th} syllable in a set of N syllables in the either expert's or learner's pronunciation.
- Using the syllable level scores, compute the score for entire stimuli as: $\frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{min(S_E(i) - S_L(i), 1)}{S_E(i)}\right)$.

Intonation

- **Why to learn Intonation:** Intonation often adds meaning to words and word group in spoken English communication [1].

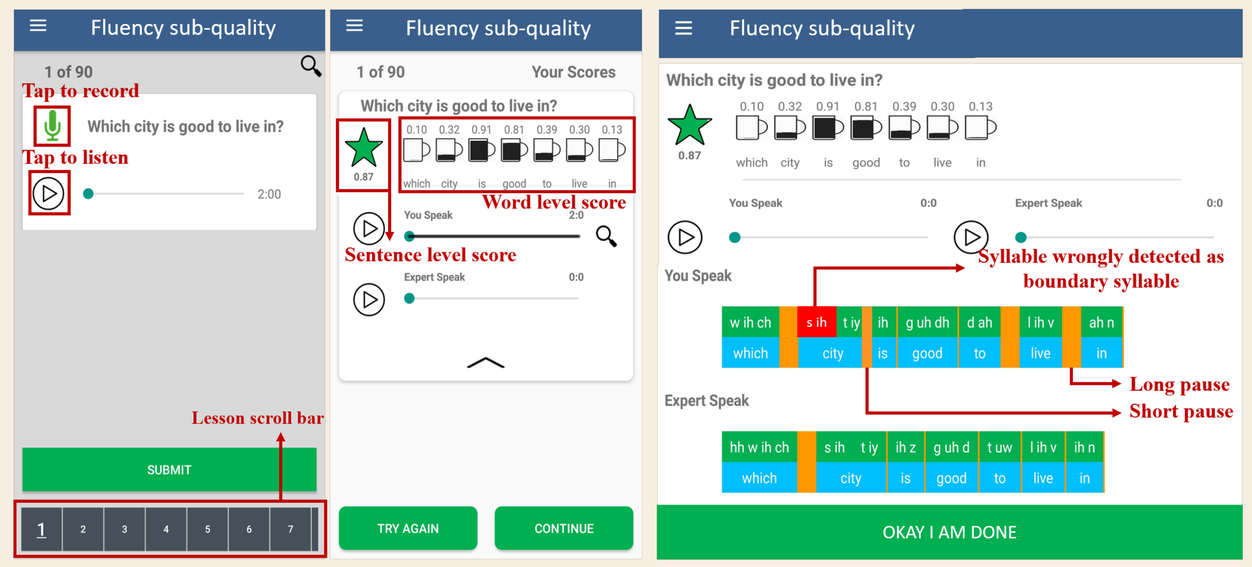


(a) Recording Interface (b) Preliminary feedback (c) Detailed feedback

- **Pitch pattern extraction:** Stylize the pitch variations in each syllable approximated with a line by minimizing mean absolute error. m and c are slope and abscissa of each line
- Perform mean and range normalize the stylized pitch and consider $|m| \geq 1$ & $m > 0 \implies$ rise tone; $|m| \geq 1$ & $m < 0 \implies$ fall tone; $|m| < 1$ & $c < 0 \implies$ low tone; $|m| < 1$ & $c > 0 \implies$ high tone.
- **Pitch pattern graph construction:** Consecutive tones, when belong to same category, are grouped.
- For each group of rise (fall) tones, we construct a line joining the lower (upper) limit from the start to the upper (lower) limit at the end.
- For the group of low (high) tones, we consider the lower (upper) limit for the entire length covered by the syllables in that group.
- **Score computation:** Correlation co-efficient between time-aligned learner's and expert's pitch patterns.

Fluency

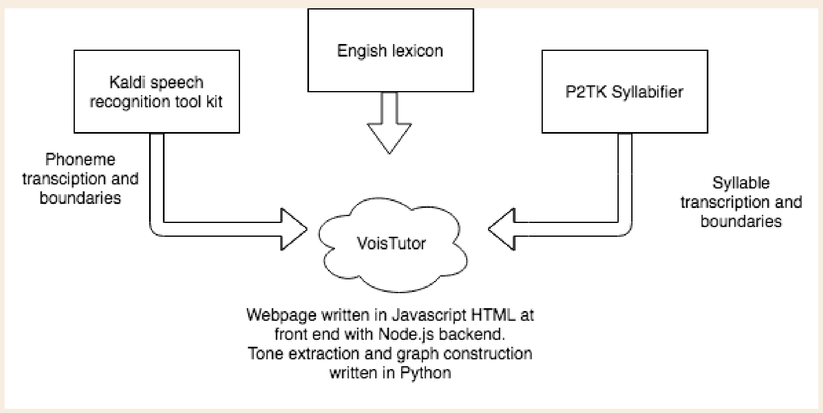
- **Why to learn fluency:** Oral fluency is considered as a measure of language proficiency and it can be improved by incorporating proper pause placement and correct pronunciation [1].



(a) Recording Interface (b) Preliminary feedback (c) Detailed feedback

- Word level score: $1 - \tanh\left(\alpha|n^E - n^L|\right)$; where, $n = \frac{\sum_{p\in w}\left(\frac{GoP(p)}{\sum_{q\in Q} GoP(q)}\right)}{N_p}$, Q is the phoneme set, p is a phoneme in word w with N_p number of phonemes, GoP is the goodness of pronunciation [3].
- Pauses are identified based on [4] and classified as long or short.
- Pause based score: Probability that a pause belongs to the same class as that of the corresponding pause in the expert's utterance.
- Sentence level score: average of word level and pause based scores.

Demostration



- The stimuli are taken from a spoken English training material [1].

Conclusion

- We present an Android app, named voisTUTOR, for improving pronunciation skills of L2 learners.
- Feedback is provided in four categories.
- We design the front end with Android SDK and back end codes with Python programming language.
- The app provides a feedback that helps for correct pronunciation of phonemes and placement of stress, intonation and pauses.
- Further investigations are required to measure the effectiveness of the proposed tool as well as analyze sufficiency of the feedback parameters on its users.

References

1. J. D. OConnor, Better English Pronunciation. (**1980**). 'Better English Pronunciation', Cambridge University Press
2. 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, pp. 5845–5849, 2017.
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. 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.

ACKNOWLEDGEMENT: Authors would like to thank the **Department of Science & Technology, Government of India** and the **Pratiksha Trust, IISc, Bangalore, India** for their support