Computer-Assisted Language Learning System: Automatic Speech Evaluation for Children Learning Malay and Tamil

Ke Shi, Kye Min Tan, Richeng Duan, Siti Umairah Md Salleh,
Nur Farah Ain Binte Suhaimi, Rajan s/o Vellu, Thai Ngoc Thuy Huong Helen, Nancy F. Chen

Institute for Infocomm Research, A*STAR, Singapore
{Shi_Ke, Tan_Kye_Min, Duan_Richeng, nfychen}@i2r.a-star.edu.sg

Abstract

We present a computer-assisted language learning system that automatically evaluates the pronunciation and fluency of spoken Malay and Tamil. Our system consists of a server and a user-facing Android application, where the server is responsible for speech-to-text alignment as well as pronunciation and fluency scoring. We describe our system architecture and discuss the technical challenges associated with low resource languages. To the best of our knowledge, this work is the first pronunciation and fluency scoring system for Malay and Tamil.

1. Introduction

Children nowadays are often encouraged to learn a second language due to globalization and national language policies. In Singapore's education system, children are required to learn their mother tongue as a second language in addition to English. The most common languages taught in schools include Mandarin Chinese, Malay, and Tamil, which correspond to the three largest ethnic groups in Singapore.

Teaching resources remain scarce, despite the continued push for learning second languages, especially for low resource languages like Malay and Tamil. Most online automatic assessment tools target languages such as English due to the availability of large amounts of linguistic resources. To the best of our knowledge, there is no known assessment tool for Malay and Tamil. Our system is trained on Malay and Tamil adult speech and testing is done on child speech to evaluate the adaptability of our system. Our system provides including numeric scores for pronunciation and fluency, and highlight words in the read sentence that could be improved. Figure 1 shows a language learner using the speech evaluation client.

Figure 1: *A student using our online speech evaluation system.*

2. System description

The architecture of our automatic assessment system is shown in Figure 2. Our system takes in a speech utterance along with the corresponding reference text as input and returns the pronunciation and fluency scores. The subsequent steps are elaborated below:

1. Each user (client) selects a sentence in the pre-defined corpus or types a self-constructed sentence in Malay or Tamil. Then the application sends a request to the assessment server after the user has read the corresponding sentence.
2. Speech-to-text alignment and decoding:
   (a) The decoder force aligns the speech samples to the reference text.
   (b) An unconstrained phone loop decoding graph is adopted to recognize the speech samples.
3. Feature extraction: The decoder extracts the duration-related rhythm features and computes the Goodness of Pronunciation (GOP) [1] scores to assess pronunciation and fluency respectively.
4. Pronunciation and fluency scoring:
   (a) A neural network takes the rhythm features such as phone duration and pauses to predict the fluency score.
   (b) Word-level scores, which are derived from phone-level GOP scores, are fed into a neural network to evaluate the pronunciation.
   (c) Words that are recommended to be improved are determined based on the pronunciation scores, fluency scores, and a set of rules designed by experienced linguists that prioritize more commonly encountered words.

The speech-to-text forced alignment and phone recognition engines are derived from the Kaldi toolkit. The GOP score for a phone p is computed as shown in the Equation (1):

$$\mathrm{GOP}(p) = \left|\log\left(\frac{p(O^{(p)}|p)}{\max_{q\in Q} p(O^{(p)}|q))}\right)\right| / \,\mathrm{NF}(p), \qquad (1)$$

where the numerator $p(O^{(p)}|p)$ is the forced alignment phone score and the denominator is the decoded phone score from the unconstrained phone loop graph. NF(p) denotes the duration of phone p. The acoustic model is implemented using a feed-forward deep neural network.

3. Scoring Framework

3.1. Pronunciation Features

Phone-level GOP scores are highly susceptible to forced alignment errors, as most terms in Equation (1) assume perfect alignment. Therefore, to present more reliable and robust assessments, we averaged phone-level scores to obtain word-level

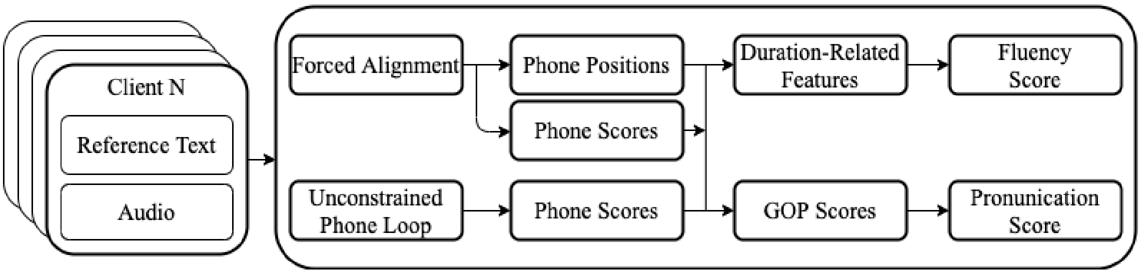
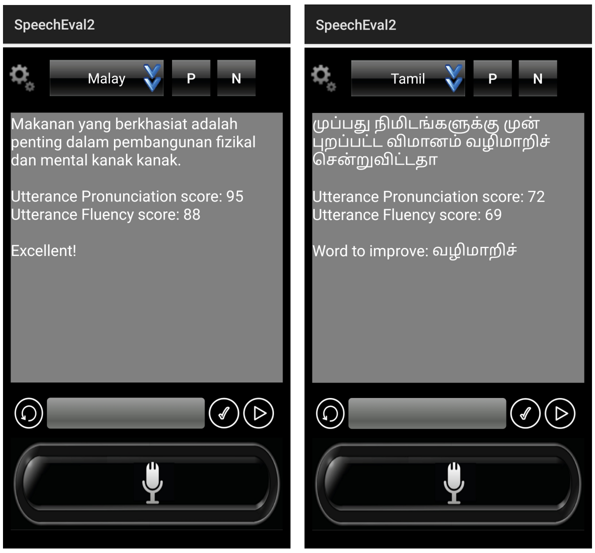


Figure 2: *System diagram of the presented speech evaluation engine*



Makanan yang berkhasiat adalah penting dalam pembangunan fizikal dan mental kanak kanak.

Utterance Pronunciation score: 95
Utterance Fluency score: 88

Excellent!

முப்பது நிமிடங்களுக்கு முன் புறப்பட்ட விமானம் வழிமாறிச் சென்றுவிட்டதா

Utterance Pronunciation score: 72
Utterance Fluency score: 69

Word to improve: வழிமாறிச்

(a) Malay evaluation example (b) Tamil evaluation example

Figure 3: *Interface showing the evaluation results.*

scores and limited pronunciation errors to the word level. Moreover, we found that word boundaries can be obtained more accurately than phone boundaries. In addition, insertion and deletion errors are more detrimental to forced alignment boundary detection than substitution errors.

3.2. Fluency Features

We define fluency as the rhythm of the read speech utterance. In particular, three types of features are adopted to evaluate fluency: (1) pause(s) within a word, (2) lengthy pause(s) between words, and (3) time duration of elongated phones.

3.3. Scoring

A feed-forward neural network is employed to evaluate the pronunciation and fluency scores based on their respective features. The outputs are then mapped to values on a scale of 0 to 100. The system considers an utterance well-pronounced or fluent based on whether the two corresponding scores meet their respective thresholds, which were determined empirically. Otherwise, the system will use the pronunciation and fluency features to determine the word(s) to be improved, and send the information to the client. Figure 3 shows the feedback results on the user interface. In particular, results in Figure 3(a) are the evaluation of a native Malay speaker with standard pronunciation and fluent rhythm, while 3(b) shows the scores of an intermediate Tamil learner's speech assessment results and the recommended word that needs more practice.

Table 1: *Scalability analysis for evaluating 7,582 utterances*

Threads	1	2	4	8	16	32
Time (min)	526	223	112	60	32	30
Memory(Mb)	753	753	753	753	753	829

4. Scalability analysis

For deployment purposes, our server needs to process multiple evaluation requests without consuming too much additional resources. 7,582 utterances were used to test our system based on a 16-core 2.70 GHz Intel Xeon CPU with 256 GB of RAM. The time needed to evaluate all utterances decreases linearly with the increase in physical CPU cores, while the memory usage only increases by an acceptable margin; see Table 1.

5. Conclusions

We summarize and share the key lessons learned in developing speech evaluation systems for low resource languages like Malay and Tamil. First, the proposed rhythm-related features are effective for fluency evaluation. Second, compared to phone-level GOP scores, word-level scores are less sensitive to alignment errors. In addition to Malay and Tamil, we are also developing Mandarin Chinese and English speech evaluation systems by taking advantage of related synergized efforts such as unsupervised adversarial multi-task training [2] and acoustic characterizations of Singaporean children's pronunciation patterns [3]. We are also working on more personalized feedback mechanisms for more effective learning outcomes.

6. Acknowledgements

This work is supported by the Agency for Science, Technology and Research (A*STAR) under its A*ccelerate Gap Fund (ETPL/18-GAP033-R20A).

7. References

[1] Silke Witt and Steve Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108, 02 2000.

[2] Richeng Duan and Nancy F. Chen. Unsupervised Feature Adaptation using Adversarial Multi-Task Training for Automatic Evaluation of Children's Speech. In *INTERSPEECH*, 2020.

[3] Yuling Gu and Nancy F. Chen. Characterization of Singaporean Children's English: Comparisons to American and British Counterparts using Archetypal Analysis. In *INTERSPEECH*, 2020.