INTERSPEECH 2021: Show & Tell Contribution
30 August – 3 September, 2021, Brno, Czechia

WittyKiddy:
Multilingual Spoken Language Learning for Kids

Ke Shi, Kye Min Tan*, Huayun Zhang, Siti Umairah Md Salleh, Shikang Ni, Nancy F. Chen*

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

Abstract

We present *WittyKiddy*, a spoken language learning system for children, developed at the Institute for Infocomm Research (I2R), A*STAR, Singapore. Our system automatically evaluates a student's oral proficiency by scoring pronunciation, fluency and intonation of a spoken utterance. We demonstrate the technical capabilities of the system via reading aloud exercises and oral cloze tests in English and Malay. Both quantitative and qualitative feedback are given to the student. Our work helps support multilingual education for children.

Index Terms: computer-assisted language learning (CALL), computer-aided pronunciation training (CAPT), speech evaluation, mispronunciation detection, Malay, children

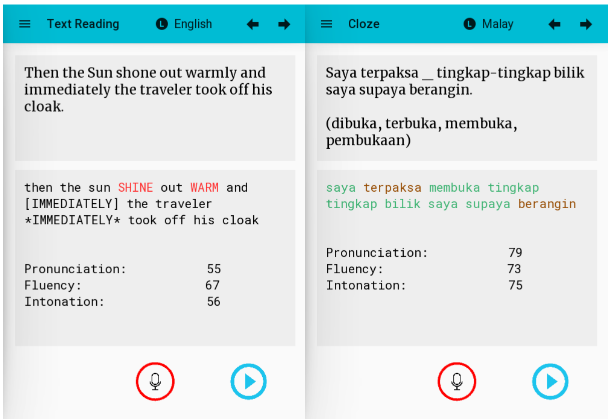
1. Introduction

In the era of globalization, children are encouraged to learn a second language. In addition, virtual learning and home-based learning have become essential given the rise of the COVID-19 pandemic. Speech evaluation aims to score speaking proficiency according to the standardized assessment criteria [1]. As speech evaluation requires one-on-one just-in-time feedback from teachers, so digital solutions to automate speech evaluation is critical to make learning and teaching more pervasive.

There are four official languages in Singapore: English, Chinese, Malay, Tamil. In addition to English, children are required to choose another language to specialize in. While speech evaluation tools for English have been well developed, automatic speech scoring on low resource languages such as Malay remains challenging due to the lack to annotated resources. As a case study, in this paper, we present a multilingual speech evaluation system that works for both spoken Malay and English. While the work is partially motivated by the multilingual demands in the Singapore context, the technology is a successful use case for other populations and countries as well.

Our system takes Goodness of Pronunciation (GOP) based on automatic speech recognition (ASR) as the backbone while incorporating fluency-aware tempo features, phonemic embeddings and pitch features. Bidirectional Long Short Term Memory (Bi-LSTM) is utilized to sequentially encode the meta features to evaluate the student's speaking proficiency.

Our system automatically detects mispronunciations and evaluates students' oral proficiency at the pronunciation, fluency and intonation levels on various tasks ranging from reading aloud text prompts to answering cloze questions or math tests orally for targeted pronunciation training. The test results on Malay and English children's speech datasets demonstrate that our speech evaluation capability is similar to human performance.



Then the Sun shone out warmly and immediately the traveler took off his cloak.

```
then the sun SHINE out WARM and
[IMMEDIATELY] the traveler
*IMMEDIATELY* took off his cloak


Pronunciation:          55
Fluency:                67
Intonation:             56
```

Saya terpaksa _ tingkap-tingkap bilik saya supaya berangin.

(dibuka, terbuka, membuka, pembukaan)

```
saya terpaksa membuka tingkap
tingkap bilik saya supaya berangin


Pronunciation:          79
Fluency:                73
Intonation:             75
```

(a) English Text Reading (b) Malay Cloze Reading

Figure 1: *The interface of WittyKiddy speech evaluation system. By clicking the record button and reading the text, the system provides fine-grained feedback on speaking proficiency. In the English example (a), the mispronunced words are highlighted: Shone → Shine is a verb tense error; warmly → warm is a parts of speech error; 'immediately' is an adverb placement error, where the read speech is not following the given text. In the Malay example (b), the word that was pronounced well is highlighted in green, while the word that needs to be improved is shown in orange. In each case, the quantitative scores on assessing pronunciation, fluency, and intonation are provided.*

2. System Description

The system consists of three parts: the acoustic model, the feature representation module and the scoring module. After a student reads a text, the speech utterance is processed by the acoustic model, where forced alignment and phone decoding is applied to obtain phoneme level timing and the acoustic log likelihood. Various features are derived to train the scoring model to predict fine-grained scores to quantify oral language skills.

Acoustic Model: The model structure is designed as a combination of different neural networks: A stack of specially designed convolution neural network (CNN) is applied at the bottom level, and a Time Delay Neural Network (TDNN) and an LSTM are placed on top of the CNN. Finally, the output of the LSTM is linked to a fully connected layer.

Feature Representation Module: Goodness of pronunciation (GOP) has been proven to be effective in speech evaluation, where phone level GOP is the time average of log posterior over the phone duration. Apart from GOP, the system also exploits tempo features , which is a combination of speaking rate and normalized phone duration. Phonemic embedding and pitch are also applied for richer and more robust representation.

*equal contribution

Scoring Module: The scoring module aims to map the meta features to a numerical score. Stacked Bi-LSTM followed by linear layer with the tanh activation function are employed to predict the scores.

3. System Features

Our system provides three types of exercises: text reading, cloze reading and arithmetic questions. Depending on the exercise, students are requested to read the provided text or answer questions aloud. With the input of the spoken utterance passed to the system, both quantitative and qualitative feedback is displayed to the user.

Mispronunciation Detection: Word and phoneme level evaluation are implemented. The pronunciation quality is reflected by the color-coded automatic transcription of the student's spoken utterance: Green → good oral delivery; Orange → words to be improved; Red → mispronunciations or errors.

language proficiency evaluation: Based on phoneme and word level acoustic/tempo analysis, 3 sentence-level holistic scores are estimated reflecting language proficiency on pronunciation, fluency, and intonation.

Grammar Error Detection: A wide variety of grammatical mistakes may be observed in the speech of non-native speakers or learners. Our system detects the most common grammar errors in pronunciation, such as: Verb Tense Errors, Subject-Verb Agreement Errors, Pronoun Errors, Double Negatives

3.1. Text Reading

Text reading is the primary function by which assessments are conducted. We included phonetically varied texts such as "The North Wind and the Sun" for the English exercise, while specially handcrafted sentences with common mistakes were prepared by linguistic researchers for the Malay exercises.

3.2. Cloze Reading

For this exercise, the student is tasked with filling in a blank in the sentence using one of several options provided. These options are selected based on their similarity in surface form or semantic representation, which may easily confuse second language learners. We use minimal pairs for English such as "sit" and "seat". For Malay, we test affixes, which alter the meaning of the root word and may lead to confusing sentences if used wrongly. For example, a few affixes of the root word "buka" along with variations are shown in Table 1.

Table 1: *Malay example of easily confused lexical choices.*

Malay Word	English Translation
buka	open
membuka	to open
dibuka	being opened
terbuka	opened (accidentally)
pembukaan	opening

3.3. Arithmetic Calculation

The last task is a number-based exercise that tests students' understanding of numerals. Students are asked to read various cardinal and ordinal numbers, as well as special formats such as years and currency. For example, the numbers "1984" should be read "one thousand nine hundred eighty-four" when describing a quantity, and "nineteen eighty-four" when it represents a year.

4. System Performance

To evaluate the system's performance, experiments are conducted in Malay and English, and the Pearson Correlation Coefficient (PCC) and Mean Square Error (MSE) between system prediction and the average score of teachers are applied as the performance metrics to evaluate system robustness.

Malay: Our Malay corpus consists of 14,088 utterances collected from 230 Singapore speakers between 9-16 years old. Utterances were scored by teachers certified by the Ministry of Education in Singapore. Each utterance is scored by teachers independently using a 5-point scale on three fronts: pronunciation, fluency, and intonation, and the inter-rater PCC were calculated between the scores of one rater and the average scores of the remaining raters. Table 2 shows that the Malay system performance approaches human performance.

Table 2: *MSE and PCC of proposed system results on Malay*

Model	Pronunciation		Fluency		Intonation	
	MSE	PCC	MSE	PCC	MSE	PCC
System	0.518	0.500	0.546	0.531	0.658	0.479
Human	-	0.547	-	0.571	-	0.545

English: The Speechocean762[1] dataset was used to evaluate the system performance in English. It consists of 5,000 English utterances collected from 250 non-native speakers while 2,500 utterances from 125 speakers are reserved as the test data. Each utterance was annotated by five raters in terms of pronunciation, fluency, and intonation, and the multiple inter-rater PCC were calculated between the scores of one rater and the median scores of the remaining raters. The English results are shown in Table 3, the PCC results suggest our system correlates well with human raters.

Table 3: *MSE and PCC results of proposed system on English*

Model	Pronunciation		Fluency		Intonation	
	MSE	PCC	MSE	PCC	MSE	PCC
System	1.368	0.654	1.037	0.702	1.064	0.699
Human	-	0.754	-	0.767	-	0.753

5. Conclusions

We presented *WittyKiddy*, a multilingual spoken language learning system for Malay and English that automatically detects mispronunciations and evaluates students' oral proficiency at the pronunciation, fluency and intonation aspects. In addition to Malay and English, we are also developing Mandarin Chinese and Tamil speech evaluation systems [2]. We are also working on more personalized feedback mechanisms for more effective learning outcomes.

6. References

[1] N. F. Chen and H. Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–7.

[2] K. Shi, K. M. Tan, R. Duan, S. U. M. Salleh, N. F. A. Suhaimi, R. Vellu, N. T. H. H. Thai, and N. F. Chen, "Computer-Assisted Language Learning System: Automatic Speech Evaluation for Children Learning Malay and Tamil," in *Proc. Interspeech 2020*, 2020.

[1] http://www.openslr.org/101/