Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Did You Mispronounce or Did I Mishear? – Detecting Kanji Mispronunciations in Children's Oral Reading

Takuya MATSUZAKI* **& Arata SAITO**
Graduate School of Science, Tokyo University of Science, Japan
*matuzaki@rs.tus.ac.jp

Abstract: Oral reading of textbooks is a common practice in Japanese elementary education, yet it poses a unique challenge: kanji characters often have multiple readings, and without a listener to provide feedback, students may unknowingly reinforce incorrect pronunciations. To address this, we propose a novel algorithm for detecting kanji mispronunciations in children's oral reading. Our method combines a deep learning-based automatic speech recognition (ASR) system with two probabilistic models: one modeling plausible kanji mispronunciations and the other modeling typical ASR errors. By aligning phoneme sequences generated from both the original text and the ASR output, the algorithm distinguishes genuine mispronunciations from transcription errors due to ASR. Experimental evaluation on speech data from children aged 6–9 shows that the proposed method successfully detects 84.6% of kanji mispronunciations that are included in the mispronunciation candidate dictionary of the probabilistic mispronunciation model.

Keywords: Kanji pronunciation, oral reading, corrective feedback, speech recognition

1. Introduction

Reading a textbook aloud is a traditional classroom activity in Japanese elementary and middle schools and still a common practice in first and foreign language classes as well as classes of other subjects. The fluency of oral reading has been identified as a key component of reading proficiency (Adams, 1990; Fuchs et al., 2001) and indicator of reading comprehension ability (Good et al., 2001; Roberts et al., 2005; Roehrig et al., 2008; Kim et al., 2010). Although there are some controversies about the directionality between reading fluency and reading comprehension (e.g., Little et al., 2017), we may at least expect that an oral reader can check her/his understanding about a text either by hearing her/him read it aloud or by feedback from a listener and may hopefully improve the understanding about the text.

Reading aloud from a textbook is thus a good candidate for a homework assignment. When doing it as homework, however, not every student has someone to listen to it. It is problematic since the execution of the homework is not guaranteed without a listener and engagement in the reading activity is naturally motivated by the existence of a listener, even if it is a robot (Nakadai et al., 2015) or a dog (Le Roux et al., 2014).

The absence of a listener causes another problem peculiar to Japanese. Kanji characters, i.e., Chinese characters in the Japanese orthographic system, generally have more than one pronunciation. For example, "上" (up) has at least the following seven pronunciations: *jou*, *shou*, *ue*, *uwa*, *kami*, *a* (as in "上げる," *a-ge-ru*), and *nobo* (as in "上る," *nobo-ru*). An oral reader has to select one of them according to the word that includes the kanji and sometimes by considering the context around the word. As a result, Japanese children, and even adults, often mispronounce a kanji in reading a text aloud. However, without a listener's corrective feedback, it is difficult for a reader to be aware of such mispronunciation and it may even strengthen the wrong association between a pronunciation and a word.

In this paper, we propose an algorithm for detecting mispronunciation of kanji in oral reading by children. Our long-term goal is to develop an AI-powered reading assistant that provides corrective feedback to children about the accuracy and fluency of their reading and motivates children's engagement in the oral reading homework. The kanji mispronunciation detection technique will be a key component of the system that helps foster the ability of choosing a proper pronunciation for kanji, which is indispensable to read Japanese text fluently.

Our kanji mispronunciation detection algorithm owes to recent advances in automatic speech recognition (ASR) based on deep learning. However, to detect mispronunciations accurately, it is not enough to simply compare the ASR result, i.e., an automatically generated transcript, against the text that was read aloud. This is because of the inherent difficulty of ASR. For instance, the word error rate (WER) of Whisper-medium model (Radford et al., 2023), which we used in the experiment, is reported to be 10.5% on a Japanese speech dataset. It means that one out of ten words is somehow misrecognized by ASR. Hence, if we report all discrepancy between the transcript and original text as potential mispronunciation, it yields too many false negatives.

To address this problem, we combine two probabilistic models: one for mispronunciation generation and the other for ASR errors. The mispronunciation generation model quantifies the plausibility of a wrong pronunciation for a word including specific kanji characters. The ASR error model captures the tendency of a phoneme to be recognized as another phoneme. By integrating them with a speech-to-text alignment algorithm, we discriminate between genuine mispronunciations and wrongly transcribed pronunciations. We evaluated the algorithm on oral reading of textbooks by children aged from 6 to 9. Experimental results show that our algorithm detects 84.6% of kanji mispronunciations included in the candidate dictionary of the probabilistic mispronunciation model.

The rest of the paper is organized as follows. Section 2 provides a summary of related work. Section 3 details the kanji mispronunciation detection algorithm. Section 4 presents experimental results. Section 5 concludes the paper.

2. Related Work

Several attempts have been made to measure the oral reading fluency and/or accuracy of students utilizing ASR (see, e.g., (Nese & Kamata, 2021) and references therein). Some recent works pursue the same goal while taking advantage of the progress in ASR by deep learning (Yıldız et al., 2024; Vaidya et al., 2024; Henkel et al., 2025; da Silva et al., 2025). While such automated assessment involves detection of mispronunciations in nature, it is not necessarily within their scope to accurately identify where and how a pronunciation error occurs in students' speech. This is because their goal is to achieve a high correlation between the fluency score obtained by ASR-based automatic method and human raters' scoring.

Another line of research has aimed at developing an automatic reading tutor based on ASR (Mostow et al., 1993; Duchateau et al., 2009; Bai et al., 2021). These systems provide feedback about missing word and pronunciation errors in oral reading in addition to the assessment of reading fluency. The main targets of these systems have however been English and other European languages. We are not aware of any previous studies that addressed the issue of identifying pronunciation errors due to logographic nature of a script system such as Chinese and Japanese, especially under the existence of ASR errors.

3. Method

Figure 1 shows an overview of our mispronunciation detection algorithm. Suppose a user read aloud a Japanese text "彼は約束を反故にした" (*kare wa yakusoku o hogo ni shita*; He broke his promise) and mispronounced the word "反故" (*ho-go*; break one's promise) as "*han ko*" because "反" may be read as "*han*" as in "反対" (*han-tai*; oppose) and "故" may be read as

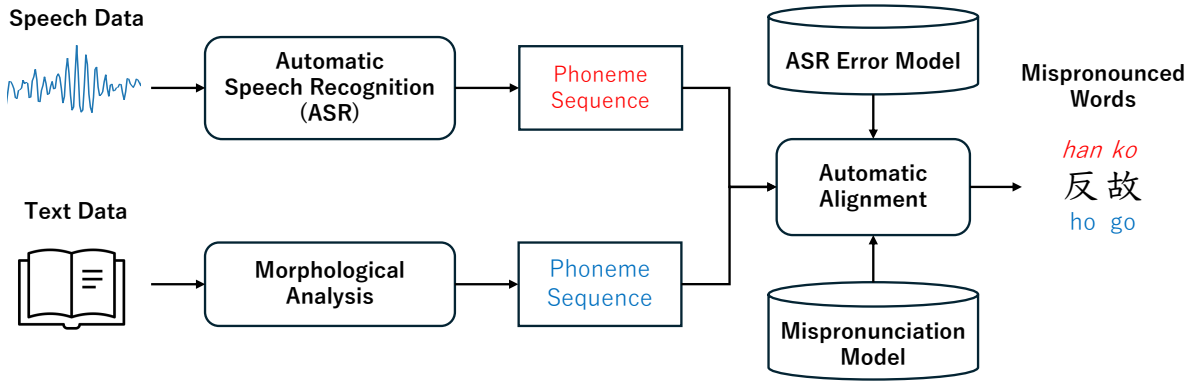

Figure 1. Overview of Mispronunciation Detection Algorithm.

"ko" as in "故郷" (*ko-kyo*; hometown). The speech data is then processed with an automatic speech recognition (ASR) system. Suppose the ASR system made a recognition error at the word "彼" (*kare*; he) and produced "*kore*" as its transcript but recognized other part of the speech correctly. We thus have the following phoneme sequence as the result of ASR:

k o r e w a y a k u s o k u o h a n k o n i sh i t a

Meanwhile, the same text is processed with a morphological analyzer. Suppose it produced a correct phoneme sequence as its reading:

k a r e w a y a k u s o k u o h o g o n i sh i t a

By comparing these two phoneme sequences, we want to detect the user's mispronunciation of the word "反故" (*han ko* / ho go) while silently ignoring the mis-recognition of the word "彼" (*kore* / kare) since it is due to an error of ASR. To this end, we utilize two probabilistic models: one for mispronunciations and the other for ASR errors. The mispronunciation model provides the (estimated) probability of the wrong pronunciations, P(*han* | 反) and P(*ko* | 故), while the ASR error model provides the probability of the mis-recognition, P(*o* | a), for "彼" (*kore* / kare). By combining these probabilities and finding the most plausible alignment between the two phoneme sequences, the system reports the mispronunciation of "反故" while ignoring the ASR error at "彼". The rest of this section explains the details of the algorithm.

3.1 Fine-tuning and Adaptation of ASR model

We used Whisper-medium (Radford et al., 2023) model for the ASR. Whisper is a neural ASR model trained on 680,000 hours of multilingual speech data collected from the internet. We further trained it for two purposes: to produce phoneme sequence instead of orthographic text (i.e., one including kanji and other Japanese characters) and to adapt it to children's speech.

To produce phoneme sequences as the ASR results, we fine-tuned Whisper using approximately 100 hours of speech with transcripts taken from the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003). The CSJ corpus is a collection of spontaneous speech by adults, such as conference talks.

The fine-tuned model's accuracy (phoneme recall) was 95.5% on adults' speech but it dropped to 90.0% on children's speech. To obtain better ASR results on children's speech, we further trained the model using 100.5 hours of speech data by 160 children (age 6 – age 8) taken from ATR's Japanese Elementary School Children Speech Database.[1] The adapted model recognized children's speech with the accuracy of 95.9%.

[1] https://www.atr-p.com/products/dbpdf/ATR_JuniorDB.pdf

3.2 Speech Recognition Error Model

To discern a mispronunciation from ASR errors, we need to quantify the plausibility of each ASR error type. Specifically, we estimated the following three types of probabilities by counting our ASR model's errors made on ATR's children speech database:

- $P(\text{ASR} = q \mid \text{pron} = p)$: the probability of outputting, either correctly or wrongly, a phoneme q for a pronounced phoneme p.
- $P(\text{ASR} = q \mid \text{pron} = \phi)$: the probability of outputting a phoneme q despite no corresponding phoneme was pronounced (i.e., phoneme q is inserted to the output).
- $P(\text{ASR} = \phi \mid \text{pron} = p)$: the probability of outputting nothing for a pronounced phoneme p (i.e., phoneme p is missing in the output).

3.3 Mispronunciation Model

Our probabilistic mispronunciation model has two components:

- $P(\text{mp} = 1 \mid \text{char} = c)$: the probability that a kanji character c is mispronounced by a child.
- $P(\text{pron} = p_1 p_2 \cdots p_n \mid \text{char} = c, \text{mp} = 1)$: the conditional probability that a kanji c is pronounced as $p_1 p_2 \cdots p_n$ given that c is mispronounced.

By estimating these two types of probabilities, we can quantify, for instance, the probability of the mispronunciation of "反故" as *han ko* as follows:

$$P(\text{mp} = 1 \mid \text{char} = 反)\, P(\text{pron} = han \mid \text{mp} = 1, \text{char} = 反)$$
$$\times\ P(\text{mp} = 1 \mid \text{char} = 故)\, P(\text{pron} = ko \mid \text{mp} = 1, \text{char} = 故).$$

Our current estimation of the probability $P(\text{mp} = 1 \mid \text{char} = c)$ is rather crude. We simply assume it does not depend on the pronounced character c:

$$P(\text{mp} = 1 \mid \text{char} = c) \ = \ \varepsilon.$$

The small constant probability ε was set to 0.05. Once we obtain a larger dataset of children's speech, we may be able to obtain a better estimate of this probability, which depends on the pronounced character as well as the traits of the reader (e.g., her/his age).

To estimate the probability of mispronouncing a kanji character c as phoneme sequence $p_1 p_2 \cdots p_n$, we counted the relative frequency of that reading in newspaper articles. Our assumption is that children tend to mispronounce a kanji character with a pronunciation that she/he hears often for that character. Since it is difficult to collect children's daily oral communication data, we instead use newspaper articles as a rough approximation of it. Specifically, we processed all the Mainichi Newspaper articles published between 1991 to 2018 by a morphological analyzer and estimated the pronunciation of each occurrence of kanji characters in the articles. By counting the number of occurrences of character c and its pronunciation $p_1 p_2 \cdots p_n$, we estimated the mispronunciation probability as follows:

$$P(\text{pron} = p_1 p_2 \cdots p_n \mid \text{char} = c, \text{mp} = 1) = \frac{\text{Count}(c \text{ with pronunciation } \ p_1 p_2 \cdots p_n)}{\text{Count}(c)}.$$

In addition to the mispronunciation probabilities, we estimated the probability that a phoneme p appearing in a text is not pronounced, denoted $P(\text{pron} = \phi \mid \text{phoneme} = p)$, and the probability that a phoneme p not appearing in a text is pronounced, denoted $P(\text{pron} = p \mid \text{phoneme} = \phi)$. We obtained these estimates by comparing the transcript and the source text in ATR's children speech database.

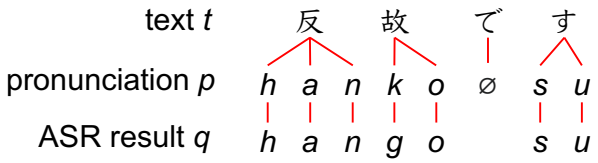

Figure 2. Alignment among Text, Pronunciation, and ASR Result.

3.4 Alignment between Speech and Text

Based on the probability models of ASR errors and children's mispronunciations, we model the generation process of children's oral reading and its speech recognition result as follows:

1. A child reads a text aloud. In reading it, she/he may mispronounce some of the kanji characters in the text, insert some unnecessary phonemes, and/or skip some phonemes, with the probabilities according to the mispronunciation model.

2. Given the recorded oral reading as input, an ASR system produces its transcript as a phoneme sequence. During it, some of the phonemes may be mis-recognized or totally missed (skipped), and some unnecessary phonemes may be inserted to the transcript, with the probabilities according to the ASR error model.

Let t denote the text that is read, $p_1 p_2 \cdots p_n$ denote the child's pronunciation of the text expressed as a phoneme sequence, and $q_1 q_2 \cdots q_m$ denote the result of ASR. The text t and the ASR result $q_1 q_2 \cdots q_m$ are observable as symbol sequences, but the true phoneme sequence $p_1 p_2 \cdots p_n$ pronounced by the child is only recorded as audio data and not directly observable as a symbol sequence. Our aim is to recover the child's true pronunciation from the observable data. We do so by finding the most probable phoneme sequence $\hat{p}_1 \hat{p}_2 \cdots \hat{p}_n$ given the text t and the ASR result $q_1 q_2 \cdots q_m$:

$$\begin{aligned} \hat{p}_1 \hat{p}_2 \ldots \hat{p}_n &= \underset{p_1 p_2 \cdots p_{n'}}{\operatorname{argmax}} P(\text{pron} = p_1 p_2 \cdots p_{n'} \mid \text{text} = t, \text{ASR} = q_1 q_2 \cdots q_m) \\ &= \underset{p_1 p_2 \cdots p_{n'}}{\operatorname{argmax}} P(\text{pron} = p_1 p_2 \cdots p_{n'}, \text{ASR} = q_1 q_2 \cdots q_m \mid \text{text} = t)\,. \end{aligned} \tag{1}$$

The mispronunciation model and the ASR error model assume an alignment among the characters in the text, the pronounced phonemes, and the phonemes in an ASR result. We denote an alignment by a. Figure 2 shows, by red lines, an example of alignment among text $t =$ "反故です" (*ho-go de su*; the promise was broken), pronunciation $p =$ "h a n k o s u", and ASR result $q =$ "h a n g o s u". Since the alignment is not observable, the rightmost-hand side of Eq. (1) takes the form of a summation over all possible alignments, but we approximate it with the most plausible alignment (i.e., so-called Viterbi-approximation):

$$\begin{aligned} \hat{p}_1 \hat{p}_2 \ldots \hat{p}_n &= \underset{p_1 p_2 \cdots p_{n'}}{\operatorname{argmax}} \sum_a P(\text{pron} = p_1 p_2 \cdots p_{n'}, \text{ASR} = q_1 q_2 \cdots q_m, \text{align} = a \mid \text{text} = t) \\ &\approx \underset{p_1 p_2 \cdots p_{n'}}{\operatorname{argmax}} \max_a P(\text{pron} = p_1 p_2 \cdots p_{n'}, \text{ASR} = q_1 q_2 \cdots q_m, \text{align} = a \mid \text{text} = t)\,. \end{aligned} \tag{2}$$

Assuming the probabilistic independence between the pronunciation of the characters in the text, we decompose the last probability as follows:

$$\begin{aligned} &P(\text{pron} = p_1 p_2 \cdots p_{n'}, \text{ASR} = q_1 q_2 \cdots q_m, \text{align} = a \mid \text{text} = t) \\ &= \prod_{i=1}^{N} P(\text{pron} = p_1^{(i)} p_2^{(i)} \cdots p_{n_i}^{(i)}, \text{ASR} = q_1^{(i)} q_2^{(i)} \cdots q_{m_i'}^{(i)}, \text{align} = a^{(i)} \mid \text{char} = c_i) \end{aligned}$$

$$= \prod_{i=1}^{N} \left\{ \begin{array}{l} P\left(\mathrm{ASR} = q_1^{(i)} \cdots q_{m_i}^{(i)} \mid \mathrm{pron} = p_1^{(i)} \cdots p_{n_i}^{(i)}, \mathrm{align} = a^{(i)}\right) \\ \quad \times P(\mathrm{pron} = p_1^{(i)} \cdots p_{n_i}^{(i)}, \mathrm{align} = a^{(i)} \mid \mathrm{char} = c_i) \end{array} \right\}, \quad (3)$$

where we assume the text t is $c_1 c_2 \cdots c_N$ as a character sequence, $p_1 p_2 \cdots p_{n'} = p_1^{(1)} \cdots p_{n_N}^{(N)}$, $q_1 q_2 \cdots q_m = q_1^{(1)} \cdots q_{m_N}^{(N)}$, and $a = a^{(1)} \cdots a^{(N)}$. The two sorts of probabilities, $P(\mathrm{ASR} \mid \mathrm{pron}, \mathrm{align})$ and $P(\mathrm{pron}, \mathrm{align} \mid \mathrm{char})$, in the rightmost-hand side of Eq. (3) can further be decomposed using the ASR error model and the mispronunciation model's component probabilities.

For instance, suppose that, as shown in Figure 2, the text is "反故です" and the reader mispronounced the first character "反" as *han* and skip pronouncing "で" (*de*). Also suppose that the ASR system misrecognized the (correct) pronunciation for "故" (*go*) as *ko*. The probability of this process is calculated as follows:

$$\begin{aligned} & P(\mathrm{pron} = \mathrm{h\ a\ n\ g\ o\ s\ u}, \mathrm{ASR} = \mathrm{h\ a\ n\ k\ o\ s\ u}, \mathrm{align} = a \mid \mathrm{text} = \text{反故です}) \\ & = P(\mathrm{ASR} = \mathrm{h\ a\ n} \mid \mathrm{pron} = \mathrm{h\ a\ n})\, P(\mathrm{pron} = \mathrm{h\ a\ n} \mid \mathrm{mp} = 1, \mathrm{char} = \text{反})\, P(\mathrm{mp} = 1 \mid \mathrm{char} = \text{反}) \\ & \times P(\mathrm{ASR} = \mathrm{k\ o} \mid \mathrm{pron} = \mathrm{g\ o})\, P(\mathrm{mp} = 0 \mid \mathrm{char} = \text{故}) \\ & \times P(\mathrm{pron} = \phi \mid \mathrm{char} = \text{で}) \\ & \times P(\mathrm{ASR} = \mathrm{s\ u} \mid \mathrm{pron} = \mathrm{s\ u})\, P(\mathrm{mp} = 0 \mid \mathrm{char} = \text{す}). \end{aligned}$$

By further decomposing the ASR recognition probabilities such like

$$P(\mathrm{ASR} = \mathrm{k\ o} \mid \mathrm{pron} = \mathrm{g\ o}) = P(\mathrm{ASR} = \mathrm{k} \mid \mathrm{pron} = \mathrm{g}) P(\mathrm{ASR} = \mathrm{o} \mid \mathrm{pron} = \mathrm{o}),$$

and substituting the estimated probabilities according to the component models, we have a probability score for the above hypothetical process.

The number of possible pronunciations $p_1 p_2 \cdots p_{n'}$ and alignments a considered in Eq. (1) is exponentially large in the length of the text. It is however possible to find the most probable pronunciation efficiently by using a dynamic programming algorithm that extends minimal edit distance calculation between two sequences (Kaji, 2023).

4. Experiment

4.1 Evaluation Data

For the evaluation of the proposed method, we recorded oral reading of paragraphs taken from textbooks by 249 elementary school students of grades 1 through 3 (ages 6 through 9). Some students provided more than one recorded file, and 370 files (6.8 hours in total) were collected. The audio speech files were manually transcribed and annotated with alignment to the text as well as mispronunciations, fillers, and repetition in the speech. The audio data and the transcription were processed with Julius (Lee et al., 2001) to obtain phoneme-level timestamps (i.e., forced-alignments) and segmented with pauses into utterances shorter than 30 seconds, which is the maximum duration of Whisper's input. On some of the speech data, the forced-alignment by Julius was failed presumably due to high noise-level of the recordings. The evaluation hereafter is based on 122 utterances (2 hours in total) thus obtained.

4.2 Mispronunciation Detection Results and Error Analysis

There were 13 occurrences of kanji mispronunciation in the evaluation data. The proposed method successfully detected 11 cases (84%) of them. Table 1 shows the mispronounced words, the speaker's mispronunciations, and the number of correctly detected ones.

Meanwhile, there were 35 cases where the system wrongly detected mispronunciation despite the reader correctly pronounced the text. We analyzed the causes of the misdetections

by inspecting the ASR output and the probability of mispronunciation and ASR errors according to the models. Table 2 provides the summary of the reason for the misdetections.

The most frequent cause of misdetection was that a word (or a part of a word) skipped by the speaker tend to be aligned with a mispronunciation that is shorter than the correct pronunciation. In our mispronunciation model, a phoneme p is assumed to be skipped by a speaker with probability $P(\text{pron} = \phi \mid \text{phoneme} = p)$, independently of its context. Thus, the probability for a word being skipped is exponentially small in the length of the word and often becomes smaller than the product of *P*(a word is mispronounced as a shorter word) and *P*(the mispronunciation is skipped). In actuality, a whole word, rather than a phoneme in it, is often skipped, and the true probability of such a skip of a whole word is much larger than one assumed in the current model. We expect that this type of error can be remedied by refining the mispronunciation model and relaxing the false independence assumption.

The second frequent cause of misdetection was due to ASR errors where the output of ASR is very close to or identical to a mispronunciation of a word despite that the speaker correctly pronounces it. In other words, a detection error of this type happens when the probability of ASR error is under-estimated. Upon inspection, we found that such problematic ASR errors often include phoneme-level errors typically made on young children's speech, such as a misrecognition between /j i/ and /ch i/. Thus, this type of error can be remedied either by a better adaptation of the ASR model to children's speech or by a refinement of the ASR error model, possibly by incorporating a speaker's trait, such as age, into the input.

Table 1. *Kanji Mispronunciations in Evaluation Data and Detection Results*

Word	Correct Pronunciation	Speaker's Pronunciation	Occurrences	Correctly Detected
大かい (athletic meet)	t a i k a i	o o k a i	1	1
金 (gold)	k i n	k a n e	1	1
せん手 (athlete)	s e n sy u	s e n t e	1	1
国 (nation)	k u n i	k o k u	1	1
年 (year)	t o sh i	n e n	1	1
大そう (very)	t a i s o o	o o s o o	1	1
見回した (look around)	m i m a w a sh i t a	m i k a i sh i t a	1	1
上あご (upper jaw)	u w a a g o	u e a g o	3	1
入れて (put in)	i r e t e	h a i cl t e	3	3

Table 2. *Reasons for Misdetections of Kanji Mispronunciation*

Cause of Misdetection	Number of Misdetections
Speaker skipped (a part of) a word, and system detected it as skipping (a part of) a mispronounced word	18
ASR system wrongly outputs a phoneme sequence that is close or identical to a mispronunciation of a word	14
Speaker repeated a word, and system detected it as a mispronunciation of a word nearby	3

5. Conclusion

This study introduced a method for detecting kanji mispronunciations in children's oral reading by integrating a fine-tuned ASR system with probabilistic models for both pronunciation errors and ASR recognition errors. The experimental results, achieving 84.6% detection accuracy, demonstrated the potential of this approach in supporting automated reading feedback. However, the analysis also revealed limitations, such as model sensitivity to phoneme skipping and underestimation of specific ASR error types. Addressing these issues through improved modeling of pronunciation context and speaker traits—such as age—could further enhance

detection accuracy. Ultimately, our method provides a core component for AI-based reading support tools aimed at improving children's reading fluency and comprehension of Japanese texts through individualized, automated corrective feedback.

Acknowledgements

This study was supported by JSPS KAKENHI Grant Number JP21H04416.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. MIT Press.

Bai, Y., Tejedor-García, C., Hubers, F., Cucchiarini, C., & Strik, H. (2021). An ASR-Based Tutor for Learning to Read: How to Optimize Feedback to First Graders. *Proceedings of Speech and Computer: 23rd International Conference (SPECOM 2021)* (pp. 58–69), Springer-Verlag, Berlin. https://doi.org/10.1007/978-3-030-87802-3_6

Duchateau, J., Kong, Y. O., Cleuren, L., Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Verhelst, W., & Van hamme, H. (2009). Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. *Speech Communication*, 51(10), 985-994.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5(3), 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Good, R. H. III, Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utilit y of a continuum of fluency-based indicators of foundational reading skills for third-grade high-sta kes outcomes. *Scientific Studies of Reading*, 5(3), 257–288. https://doi.org/10.1207/S1532799XS SR0503_4

Henkel, O., Horne-Robinson, H., Hills, L., Roberts, B., & McGrane, J. (2025). Supporting Literacy Assessment in West Africa: Using State-of-the-Art Speech Models to Assess Oral Reading Fluency. *International Journal of Artificial Intelligence in Education*, 35, 282–303. https://doi.org/10.1007/s40593-024-00435-9

Kaji, N. (2023). Lattice path edit distance: A Romanization-aware edit distance for extracting misspelling-correction pairs from Japanese search query logs. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 233–242). Association for Computational Linguistics.

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, 102(3), 652–667. https://doi.org/10.1037/a0019643

Lee, A., Kawahara, T., & Shikano, K. (2001) Julius – An Open Source Real-Time Large Vocabulary Recognition Engine. *Proceedings of 7th European Conference on Speech Communication and Technology* (pp.1691–1694). International Speech Communication Association.

Little, C. W., Hart, S. A., Quinn, J. M., Tucker-Drob, E. M., Taylor, J., & Schatschneider, C. (2017). Exploring the co-development of reading fluency and reading comprehension: A twin study. *Child Development*, 88(3), 934–945. https://doi.org/10.1111/cdev.12670

Maekawa, K. (2003). Corpus of Spontaneous Japanese: its design and evaluation. *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (pp. 7-12).

Mostow, J., Hauptmann, A. G., Chase, L. L., & Roth, S. (1993). Towards a reading coach that listens: automated detection of oral reading errors. *Proceedings of the eleventh national conference on Artificial intelligence* (pp. 392–397). AAAI Press.

Nakadai, H., Hee, L. S., Kitajima, M., Hoshino, J. (2015). KINJIRO: Animatronics for Children's Reading Aloud Training. *Entertainment Computing - ICEC 2015* (pp. 252-260). Springer-Verlag. https://doi.org/10.1007/978-3-319-24589-8_19

Nese, J. F. T., & Kamata, A. (2021). Evidence for automated scoring and shorter passages of CBM-R in early elementary school. *School Psychology*, 36(1), 47–59. https://doi.org/10.1037/spq0000415

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the International Conference on Machine Learning*, (pp. 28492-28518). JMLR.org.

Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading compr ehension. *School Psychology Quarterly*, 20(3), 304–317. https://doi.org/10.1521/scpq.2005.20.3. 304

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, 46(3), 343–366. https://doi.org/10.1016/j.jsp.2007.06.006

Le Roux, M. C., Swartz, L., & Swart, E. (2014). The effect of an animal-assisted reading program on the reading rate, accuracy and comprehension of grade 3 students: A randomized control study. *Child & Youth Care Forum*, 43(6), 655-673.

da Silva, G. C., Rodrigues, R. L., Amorim, A. N., Jeon, L., Albuquerque, E. X. S., Silva, V. C., da Silva, V. F., Pinheiro, A. L. A., Nunes, J. P. J. R., de Souza, S. X. M. G., Silva, M. S., Mauro, I., & Maciel, A. M. A. (2025). Assessing reading fluency in elementary grades: A machine learning approach, *Computers and Education: Artificial Intelligence*, 8, 100411.

Vaidya, M., Sahoo, B. K., & Rao, P. (2024). Deep Learning for Assessment of Oral Reading Fluency. arXiv. https://doi.org/10.48550/arXiv.2405.19426

Yıldız, M., Keskin, H. K., Oyucu, S., Hartman, D. K., Temur, M., & Aydoğmuş, M. (2024). Can Artificial Intelligence Identify Reading Fluency and Level? Comparison of Human and Machine Performance. *Reading & Writing Quarterly*, 41(1), 66–83.