WOCCI 2017: 6th International Workshop on Child Computer Interaction
13 November 2017, Glasgow, Scotland, UK

Developing speech processing technologies for shared book reading with a computer

Anastassia Loukina, Beata Beigman Klebanov, Patrick Lange, Binod Gyawali, Yao Qian

Educational Testing Service, USA
{aloukina, bbeigmanklebanov, plange, bgyawali, yqian}@ets.org

Abstract

We present a preliminary report on developing technology for an application that supports shared book reading. We discuss how speech processing technology can be used to automate different components of the system for oral reading fluency evaluation during shared book reading and the challenges posed by this new context in comparison to other automated reading tutor systems. We also present performance evaluation of the baseline system on a corpus of read speech.

Index Terms: reading fluency assessment, child speech, automated speech recognition

1. Introduction

The importance of knowing how to read is hard to overestimate. Accordingly, reading is a cornerstone of the K-12 education in the United States: Along with mathematics, a biennial national assessment of reading skills in grades 4th and 8th is required by law[1] in order to inform education policy. According to the 2015 report, 31% of U.S. 4th graders read below the Basic level.[2] The goal of our research is to use technology to support the development of reading fluency in children.

Shared oral reading is a beloved activity that starts as early as infancy in many families. According to the Kids & Family Reading Report, a national survey of children ages 6-17 and their parents exploring attitudes and behaviors around books and reading in the U.S. published by Scholastic in 2016,[3] 40% of parents read aloud to their children before they were 3 months old, and 62% reported doing so almost every day with their 3-5 year old kids. 87% of children aged 6-11 who were read to aloud at home report liking or having liked the activity.

While the frequency of being read aloud to drops for children over the age of 5, other varieties of shared reading are practiced well into the elementary school [1, 2, 3, 4], often as a method to help weaker readers, such as dyad reading with a more fluent peer [5, 6, 7], or reading with an adult volunteer or parent [8, 9, 10]. Notably, when reading together with a higher-proficiency partner, the weaker reader has access to more complex, and thus potentially more engaging, reading materials than the weaker reader could have read independently. Thus [5] experimented with dyads reading texts 2, 3, and 4 grade levels above the instructional level of the assisted reader; results showed robust gains in oral reading fluency and comprehension in assisted readers across conditions.

In this paper, we explore the speech technology necessary to support the virtual reading companion described in [11]. This system uses the recording of an expert adult reader as a reading partner that takes turns reading a book with the child and speech processing technologies to process child's speech. We selected Harry Potter and the Sorcerer's Stone (HP1) by J. K. Rowling as the book to be read, in order to maximize the potential for engaging the child with a good story. Analyzing data on children's choices of books along with comprehension quizzes taken by more than 150,000 children in the U.K., [12] concludes that "In the early grades, children are reading very difficult books with a high degree of success. The effect of reading highly motivating books is remarkable. Chief among these are the Harry Potter books." For the adult reader, we use the recorded narration by the award-winning actor Jim Dale [13].

During the narrator's turns, the child can follow along on the screen or just listen. When it is the child's turn to read, an automated speech analysis system would capture and process the child's oral reading in order to adjust the system's behavior (the narrator might read more to a weaker reader), provide feedback, or track improvement in reading fluency. While automated systems have been successfully used before for assessment of oral reading on short passages [14, 15, 16], our application poses significant additional challenges:

- Children reading varying excerpts from the whole book, not a fixed set of passages; the language model used for the automatic speech recognition (ASR) thus needs to be quite general.
- Passages are unedited excerpts from a book and thus include dialogues between characters, various cases of onomatopoiea, etc. These could elicit unusual speech patterns that would in turn affect the performance of the automated system.
- We expect our system to be used in a classroom or at home and therefore there is a high likelihood of background noise including other children reading aloud the same text. Furthermore, the goal is to encourage reading, not to assess; therefore, we do not want to penalize off-task speech. However, we want to make sure we are not analyzing off-task speech as reading for the purposes of tracking fluency.

In this paper we describe the approach we took for building the first version of the system including ASR, off-task speech detection and computation of fluency measures. We then present its evaluation on a small corpus of recordings of children-read speech collected for this project. An additional practical consideration for this version is that the speech processing component needs to be incorporated into the process for usability testing long before sufficiently many children read the whole book to generate enough in-domain data for training the system. Therefore we evaluate whether it is possible to achieve reasonable performance with a system trained entirely on external data.

[1] The Elementary and Secondary Education Reauthorization Act of 2001; see "Assessment Policy" in https://nces.ed.gov/nationsreportcard/about/assessmentsched.aspx

[2] https://www.nationsreportcard.gov/reading_math_2015/ #reading/acl?grade=4

[3] http://www.scholastic.com/readingreport/about.htm

10.21437/WOCCI.2017-8

2. Related work

2.1. Automated reading tutors

There exist many commercial and research applications designed to assess oral reading fluency and assist with its development; it would not be possible to provide a comprehensive overview in this paper. Reviews of earlier systems can be found in [17, 18, 15] among many others; [19] provide an overview of some of more recent developments in the area of technology-based literacy instruction. VersaReader [15] and Project LISTEN [14] are two of the most mature systems in this area, while the self-administered app Moby.Read [16] is the latest published newcomer. All these systems rely on ASR to process children's speech and then use the hypothesis to compute various oral fluency measures. Their findings show that this approach is very effective and results in measurements that show high agreement with those assigned by human raters.

However, the content read by the children in the majority of published applications is short and grade-controlled, often designed specifically for reading practice or for assessment of reading. In our application, the children read an original book. As shown in [11], there is a large variation in textual features across passages in this book, including variation in estimated grade level, which may affect the reading patterns and in turn the performance of the system. While a lot of research on ASR considered the impact of ASR errors on educational applications in general [20] and reading tutors in particular [14, 16], less is known about how the content of read speech might affect ASR accuracy. For conversational speech, [21] identified several major factors that affect ASR accuracy including words preceding interruption points and words with extreme prosodic characteristics. The latter are more likely to occur in a book than in a test passage and might decrease system performance.

Finally, very few papers on reading tutors consider identification of off-task speech as part of the design. Sometimes this is because the system is designed to work with a timed assessment (as in [15] where off-task speech would be unlikely and if present might be considered during scoring). An algorithm for off-task speech detection is part of the Reading Tutor from Project LISTEN [14, 22]. Since the Reading Tutor breaks the reading into short sentences, this algorithm is designed to classify each recording as either off-task or on-task. A different approach would be necessary for our application where we need to distinguish on-task and off-task speech *within* the same recording.

3. System description

3.1. Automated speech recognition

The first version of our system for shared book reading is not expected to change its behavior based on the content of any off-task speech produced by the child. Therefore our goal for the ASR engine was to be able to identify when the child is reading the book and to achieve high accuracy when recognizing the child's reading of the text.

The ASR was trained using the Kaldi toolkit [23]. This system was initially developed for recognizing read and spontaneous speech from non-native children in the context of language proficiency assessment and is described in more detail in [24]. We used the version referenced in that paper as 'DNN'. This system is based on feed-forward DNN-based speech recognizer with i-vectors for speaker adaptation. The acoustic models were trained on a corpus of almost 8,000 spoken responses to English language proficiency assessment for children between 11 and 15 years old (approx. 140 hours of speech). [24] reported an ASR word error rate (WER) of 7% on read speech for in-domain data.

The language model was trained using the text of the book only. There were 783 out-of-vocabulary words that were not in the original CMU dictionary[4] used for model training. We generated transcriptions for these words using an automatic grapheme-to-phoneme converter [24] and then corrected them manually as necessary. For these initial experiments we decided to train one language model per each chapter of the book and trained the language model on Chapter 1. Since the book chapter is relatively short, we concatenated it multiple times to artificially increase the 3-gram and 2-gram count.

We used ASR performance on the audio book to test and fine-tune the language model and the weights for language model and acoustic model. To do this we extracted from the first chapter 18 non-overlapping passages, each about 250 words in length. These were then recognized automatically and the ASR hypothesis was compared to the original book. Even though our acoustic models were optimized for children and therefore would be a mismatch for the adult male narrator, we expected that the system's performance should still remain accurate because the narrator's reading is very well enunciated and free of disfluencies or any off-task text.

The average WER across all 18 passages extracted from Chapter 1 of the audio book was 2.3%. This is consistent with our expectation that acoustic mismatch should not have detrimental effect on system performance and the WER for such clear speech is expected to be very low.

However, we also found that WER varied across 18 passages in our sample: one passage was an outlier with a WER of 12%. In this passage the narrator performed sobbing of one of the characters and that fragment was not recognized correctly. For the remaining 17 passages we saw a clear bimodal distribution with WER ranging between 0% and 1.6% (mean .8%) for 12 passages and between 2.8% and 4.6% (mean 3.8%) for the remaining 5 passages. This pattern might deserve further investigation if it persists on a larger sample of passages from the rest of the book.

3.2. Identification of off-task speech

We expect that some of the recordings might contain at least some off-task speech. In order to make sure our analyses are only based on child reading the text, we developed a baseline module for identifying such speech when it occurs before or after the on-task reading.

We used an algorithm similar to the one described in [22] where the ASR hypothesis is aligned back to the expected prompt and then identified the timestamps for the beginning of the first and the end of the last word aligned to the prompt. Note that in [22] this algorithm was applied to human transcriptions to create gold-standard for system evaluation while here we use it for actual identification of off-task speech.

3.3. Measures of oral reading fluency

We use the ASR hypothesis and the associated timestamps to compute two measures of oral reading fluency: words correct per minute and reading accuracy.

Words correct per minute (WCPM) is a standard measure of oral reading fluency which combines aspects of speed and

[4] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

accuracy and has been shown to be a good predictor of reading skills [25, 26]. It was computed as the total number of correctly read words divided by the total time it took the child to read the passage based on the timestamps we estimated for the beginning and end of on-task speech.

The second measure, reading accuracy, was computed as the total number of correctly read words divided by the total number of words in the passage text. Note that this measure penalizes deletions and substitutions but not insertions and is not dependent on speed of reading.

4. System evaluation

4.1. Corpus of read speech

We evaluated the system on a new corpus of children's oral reading collected for this project. The corpus is described in more detail in [11].

The subset of the corpus used in this study includes 66 recordings of 22 children reading 3 texts from HP1. At the time of the recording (April 2017), all children attended 2-4 grade (6-8 children per grade, 12 girls and 10 boys) and were selected via a convenience sample.

The recordings took place in an office with 2-3 children recorded simultaneously. Before reading the experimental passages, the child listened to the very first passage from the HP1 audio-book. Then the child read aloud the passage immediately following the passage read by the narrator. In the original data collection, the children read three more passages presented to them in a randomized order[5]. The children were asked to read at their natural pace. The texts were presented on a laptop screen and captured via headset microphone.

In this study we use the first passage read by the children and two of the other three passages since one of the passages was from Chapter 2. The passages contained 246, 226 and 306 words. Two of the texts contained a lot of dialogue, one text included stuttering.

On average it took children a bit over 2 minutes to read each text. The experiment was set up in a way that children could not proceed to the next text until 3 minutes from the start of the recording and the duration of most of the recordings is 3 minutes. Some children talked to the experimenter after the end of their reading turn. Some recordings also captured background noise from other children reading during the silence.

4.1.1. Off-task and on-task speech in transcriptions

We used a professional transcription agency to obtain word-by-word transcriptions of each child's reading. In addition to transcriptions, the transcribers also indicated whether speech was on-task or off-task and recoded the timestamps for each such transition. Among the 66 recordings, 19 recordings (28%) included at least some off-task speech, mainly the conversation between the child and the proctor after or sometimes before the reading. The amount of off-task speech varied from a couple of words to 110 words, with the median of 27 words. In this dataset, the off-task speech always occurred either before or after the recording, there were no instances of off-task speech in the middle of the reading.

Since timestamps are recorded in seconds, we had another set of professional linguists record the beginning and end of on-task speech to estimate the measurement error. Average difference between the two sets of stamps was 1.15s with an almost perfect correlation in estimated duration of on-task speech (r = .996).

4.1.2. Computation of gold standard oral reading measures

To compute the gold standard values for reading accuracy and WCPM, we aligned the transcriptions to the passage text and used the method described in 3.3.

Most children's reading closely followed the texts, with the average of 97.7% of all words in each text read correctly (SD = 2.2, min = 87.5%, max = 100%). The average WCPM in the 3 texts in our corpus was 116.0 (SD = 22.5, min = 59.0, max = 166.8). As also noted in [11], when compared to peers of the same age, these children generally read quite fluently and accurately: a grade-stratified sample of children from grades 2-4 during spring term is expected to read, on average, at 106 WCPM [27]. The observed rate of 117 WCPM corresponds to 60% percentile – somewhat above average. Note that this is only a rather rough estimate of these children's fluency relative to peers, since the experimental texts differ in complexity substantially from the grade-leveled materials used for oral reading fluency assessments.

4.2. System performance

4.2.1. ASR accuracy and identification of off-task speech

We first used the algorithm described in 3.2 to separate on-task and off-task speech. We aligned the ASR hypothesis to the prompt and used this alignment to automatically establish the timestamps for when the child started and stopped reading. We next evaluated how these timestamps compared to those recorded by the human transcribers across all 66 responses in our corpus. Average absolute error between both starting and final stamps was 1.5 seconds. Paired-sample t-test showed that there was no statistically significant difference in average discrepancy between the stamps assigned by two human annotators vs. average discrepancy between automated and human estimates (t = 1.50, p = .13). We also observed an almost perfect correlation in estimated duration of the reading with r = .99. For comparison, had we used the initial and final ASR timestamp for the whole recording, without excluding off-task speech, to estimate the duration of the reading, the correlation between actual and estimated duration would have gone down to r = .33.

We next computed two values of WER for each response: (a) using the ASR hypothesis and transcription for the whole recording; (b) using only the part of the hypothesis identified automatically as on-task speech and the transcription marked by the transcribers as 'on-task' speech. As expected, for 47 responses where human transcriptions included off-task speech, the WER computed for on-task speech only was lower than WER computed for the whole response including off-task speech: 10.2% vs. 22.1%. Notably, removing automatically identified off-task speech also lead to decrease in WER for 19 responses where human transcription only included on-task speech: 9.9% vs. 11.8% for the whole response. The improvement in WER for these responses was due to the fact that the ASR hypothesis in some cases contained 'host' off-task speech where the recognizer attempted to recognize another child reading in the background. Since that background reading was not part of human transcription, the overall WER for the whole recording was higher than when computed on on-task part of

[5] In actual application the children would have read the texts in the same order. In this study order randomization was done to allow separation between text and order effects in subsequent analyses. See [11] for further discussion.

the recording only.

4.2.2. *Accuracy of automated reading fluency measures*

We used the ASR hypothesis and timestamps to compute the two oral reading fluency measures discussed in 3.3 and evaluated how they compare to the same measures computed based on human transcription and timestamps. The measures were computed using only automatically identified on-task speech. We observed an almost perfect correlation for WCPM (r = .98) and a moderate correlation for reading accuracy (r = .57). To further understand the difference in performance on these two measures, we evaluated the accuracy of estimation of the total number of correctly read words, the numerator for both measures. We found that the difference between automatic estimates and those based on human transcription were on average 2.5 words or about 1% of the total number of words. Since the overall accuracy of children reading was very high with low variation across children, even relatively small errors had substantial detrimental effect on the correlation for accuracy.

We further evaluated, how much worse our estimates would be without excluding off-task speech. As expected, including off-task speech had no effect on reading accuracy since this measure does not penalize insertions. However, it had substantial effect on our estimates of WCPM with correlation going down to r = .36.

4.2.3. *Performance variation across texts*

Our analysis of ASR performance on narrator data suggested variation across texts. Therefore we further explored whether there were any text-based differences in automated system performance.

We first confirmed that the differences in oral reading fluency across texts reported in [11] persist in this sample (note that we use the data from only some of the children considered in that study and two out of three texts. We also added an additional text that was previously used as a control). That study reported statistically significant differences in WCPM across the three texts used for analysis.

Table 1 shows the WCPM and reading accuracy for the three texts in this study.[6]

Table 1: *Oral reading fluency measures computed based on human transcription for the three texts in our study*

Text	WCPM	Accuracy	WER (child)	WER (narrator)
Text 1	129.6	98.5%	8.1%	3.6%
Text 2	114.2	97.0%	9.9%	.9%
Text 3	104.2	97.6%	11.2%	12.1%

Mixed level models with WCPM as dependent variables, speaker as random effect and text ID as fixed effect showed that Text 1 had the highest WCPM and Text 3 had the lowest WCPM (p < .0001 in both cases). Thus the difference in WCPM reported in [11] persist in this subset of the corpus. Relative to WCPM, the differences between the texts were less pronounced in terms of reading accuracy: the accuracy for Text 1 was higher than for Text 2 (p = .01), but there was no significant difference between Text 1 and Text 3 (p = .10) or Text 2 and Text 3 (p = .36).

We next looked at WER for on-task speech for these texts. Using the same mixed effects model approach, we found that the WER for Text 1 was lower than Text 3 (p < .01), yet there was no statistically significant difference between Text 1 and Text 2 (p = .1) or Text 2 and Text 3 (p = .06). Note that these patterns did not align with the patterns we observe for WER computed on narrator data: for narrator, Text 2 had the lowest WER and Text 3 was the outlier with a particularly high WER discussed in more detail in 4.2.1.

However, these patterns of WER match those we observed for oral reading fluency: Text 1 which elicited most fluent reading as measured by the traditional measures also had the lowest WER. The lower WER for more accurate reading is not surprising since ASR is unlikely to accurately recognize substitutions and insertions and may generally be affected by disfluent reading [21]. Therefore we tested whether the differences in WER between texts would remain if we control for WCPM in child reading as measured by transcription. After adding WCPM based on transcription as a fixed factor to the model, we found that the differences in WER between different texts were no longer significant, but there was a strong dependency between WCPM and WER (p < .0001).

4.3. Can we make the language model more general?

For this first version of the system we trained the language model on the text from Chapter 1. Such chapter-based approach would require maintaining 17 different models. Therefore we further explored whether training the language model on the whole book would result in substantial degradation in system performance. We retrained the language model using the same approach on the full text of the book and repeated the same analyses. We found that the WER for on-task speech increased substantially from 10% to 40% with substantial variation across the three texts (25%, 53% and 41% for Texts 1, 2, 3). The correlations for automated estimates of WCPM and reading accuracy based on these ASR outputs using full-book language model also decreased to r = .78 and r = .43 respectively.

5. Discussion and directions for future work

In this paper we discussed the challenges of developing speech analysis technology for supporting shared book reading. In order to be able to include ASR into usability rounds, we used external data and the book text to train the ASR. Our system achieved WER of 10%. This is higher than the performance of original system on on-domain data (7%) but is comparable to WER reported for other systems used in automated reading assessment: in their overview of child ASR, [28] cite WER of 8-12% for different systems.

The ASR hypothesis and timestamps could be used to compute WCPM, the dominant measure of reading fluency, with almost perfect accuracy (average r across three texts = .98). At the same time the correlation for reading accuracy was lower with r = .57. There are two related reasons for this discrepancy: first of all, as discussed in 4.1.2, the overall reading accuracy in this corpus was very high with little variation across children. As a result, relatively small errors in accuracy estimation (2-3 words or 1% of the text) could have a major effect on the overall correlation. On the other hand, for these accurate readers words correct per minute is very close to words per minute (see also [11]) and thus is predominantly determined by the correct esti-

[6] In [11] Text 1 is referred to as a 'control text', Text 2 is 'Easy' text and Text 3 is 'Hard' text

mation of the overall duration of on-task speech. As reported in 4.2.1, our estimates of duration were highly accurate with r = .99.

Our results also showed that while training the language model on one chapter made it possible to accurately recognize different passages from that chapter, a language model trained using the same approach on the whole book performed much worse (WER of 40%). Note that in addition to the obvious reason that book-based language model is too broad, it could also be that our method does not work as well on longer texts. We will explore different approaches to language model training in future studies.

We also considered off-task speech which is likely to be present when speech is collected in informal context. Furthermore, as sometimes happened in our data, 'host' off-task speech can be inserted by ASR when the speaker is silent due to background noise. Very few descriptions of automated systems for oral reading fluency assessment mention identification of off-task speech as a separate task. While sometimes this may be justified, especially in the assessment context, for our application off-task speech should not be considered when estimating oral reading fluency and thus we need to be able to identify it for the measurements to remain valid. Our empirical results also show that not excluding off-task speech leads to a substantial decrease in performance. We found that a baseline system based on relatively simple string-matching algorithm could already achieve accurate performance when identifying the off-task speech before and after the on-task reading. Further research is necessary to establish the best way to approach off-task speech in the middle of the reading both in terms of identification but also in terms of how it should be treated when computing oral reading fluency measures.

One of the concerns when developing technology for shared book reading is the type of the material: an original book might contain a lot of different stylistic devices that might elicit unusual speech patterns which in turn would cause ASR failure. While our results for narrator suggested that there may be text effect on ASR performance, we did not observe these for the three texts read by children despite the fact that these three texts elicited different fluency patterns. The ASR performance was primarily driven by the fluency of children's reading with no additional effect of text beyond those already manifested in fluency patterns. A larger sample of texts would of course be necessary to confirm this finding.

The main limitation of this study is the very small size of the corpus and the fact that the children in the sample are relatively fluent readers. Even for this sample, however, we found that the ASR accuracy is strongly related to child's oral reading fluency. Therefore the system performance might be lower when evaluated on a more diverse sample.

Finally, in this pilot study we evaluated the system performance using two standard measures of oral reading fluency: words correct per minute and accuracy. Yet in addition to the ability to read quickly and accurately, the construct of oral reading fluency also includes the ability to read with a natural intonation [29, 30], an aspect not covered in this study. Furthermore, WCPM, one of the two measures we used in this study has been shown to exhibit variation across texts [25, 11, 16]. This raises questions about the validity of using this measure for continuous fluency tracking. Our results suggest that there may also be text-based variation in accuracy even though for this data the effect appears to be smaller. We will explore these questions in future research.

6. Acknowledgements

We would like to thank René Lawless, Kelsey Dreier, and Thomas Florek for their help with data collection and Jimmy Bruno, Zydrune Mladineo and Patrick Houghton for help with annotation. We would also like to thank Su-Youn Yoon, Vikram Ramanarayanan, Keelan Evanini, two anonymous WOCCI reviewers and our many colleagues at ETS who discussed this work with us and provided useful comments.

7. References

[1] C. Young, K. A. J. Mohr, and T. Rasinski, "Reading together: A successful reading fluency intervention," *Literacy Research and Instruction*, vol. 54, no. 1, pp. 67–81, 2015.

[2] J. Mansell, M. A. Evans, and L. Hamilton-Hulak, "Developmental changes in parents' use of miscue feedback during shared book reading." *Reading Research Quarterly*, vol. 40, no. 3, pp. 294–317, 2005.

[3] E. B. Meisinger, P. J. Schwanenflugel, B. A. Bradley, and S. A. Stahl, "Interaction quality during partner reading," *Journal of Literacy Research*, vol. 36, no. 2, pp. 111–140, 2004.

[4] J. L. Eldredge, D. R. Reutzel, and P. M. Hollingsworth, "Comparing the effectiveness of two oral reading practices: Round-robin reading and the shared book experience," *Journal of Literacy Research*, vol. 28, no. 2, pp. 201–225, 1996.

[5] L. T. Brown, K. A. J. Mohr, B. R. Wilcox, and T. S. Barrett, "The effects of dyad reading and text difficulty on third-graders' reading achievement," *The Jorunal of Educational Research*, vol. Advance online publication, 2017.

[6] A. Morgan, B. R. Wilcox, and J. L. Eldredge, "Increasing reading performance of low-achieving second graders with dyad reading groups," *Journal of Education Research*, vol. 94, no. 2, pp. 113–119, 2000.

[7] J. L. Eldredge and D. W. Quinn, "Increasing reading performance of low-achieving second graders with dyad reading groups," *The Journal of Educational Research*, vol. 82, no. 1, pp. 40–46, 1988.

[8] N. F. Knapp, "Reading together: A summer family reading apprenticeship program for delayed and novice readers," *Literacy Research and Instruction*, vol. 55, no. 1, pp. 48–66, 2016.

[9] ——, "Cougar readers: Piloting a library-based intervention for struggling readers," *School Libraries Worldwide*, vol. 19, no. 1, 2013.

[10] N. F. Knapp and A. P. Winsor, "A reading apprenticeship for delayed primary readers," *Reading Research and Instruction*, vol. 38, no. 1, pp. 13–29, 1998.

[11] B. Beigman Klebanov, A. Loukina, J. Sabatini, and T. O'Reilly, "Continuous fluency tracking and the challenges of varying text complexity," in *Procceings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark.: Association for Computational Linguistics, 2017, pp. 22–32.

[12] K. J. Topping, "Fiction and non-fiction reading and comprehension in preferred books," *Reading Psychology*, vol. 36, pp. 350–387, 2015.

[13] J. K. Rowling and J. Dale, "Harry Potter and the sorcerer's stone," [New York], 2016.

[14] J. Mostow, "Why and how our automated reading tutor listens," *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, pp. 43–52, 2012.

[15] J. Balogh, J. Bernstein, J. Cheng, A. Van Moere, B. Townshend, and M. Suzuki, "Validation of automated scoring of oral reading," *Educational and Psychological Measurement*, vol. 72, no. 3, pp. 435–452, 2012.

[16] J. Bernstein, J. Cheng, J. Balogh, and E. Rosenfeld, "Studies of a Self-Administered Oral Reading Assessment," in *Proceedings of SLaTE 2017*, O. Engwall, J. Lopes, and I. Leite, Eds. Stockholm: KTH Royal Institute of Technology, 2017, pp. 180–184.

[17] M. Eskenazi, “An overview of spoken language technology for education,” *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[18] K. Zechner, J. Sabatini, and L. Chen, “Automatic Scoring of Children’s Read-Aloud Text Passages and Word Lists,” *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 10–18, 2009.

[19] S. A. Crossley and D. McNamara, “Educational Technologies and Literacy Development,” in *Adaptive Educational technologies for literacy instruction*, S. A. Crossley and D. Mcnamara, Eds. New York: Routledge, 2016, pp. 1–12.

[20] J. Tao, K. Evanini, and X. Wang, “The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system,” in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 294–299.

[21] S. Goldwater, D. Jurafsky, and C. D. Manning, “Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates,” *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[22] W. Chen and J. Mostow, “A tale of two tasks: Detecting children’s off-task speech in a reading tutor,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 236, pp. 1621–1624, 2011.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[24] Y. Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland, “Bidirectional LSTM-RNN for Improving Automated Assessment Of Non-native Children’s Speech,” in *Proceedings of Interspeech 2017*. Stockholm: International Speech Communications Association, 2017, pp. 1417–1421.

[25] S. P. Ardoin, T. J. Christ, L. S. Morena, D. C. Cormier, and D. A. Klingbeil, “A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules,” *Journal of School Psychology*, vol. 51, no. 1, pp. 1–18, 2013.

[26] M. M. Wayman, T. Wallace, H. I. Wiley, R. Tich, and C. A. Espin, “Literature synthesis on curriculum-based measurement in reading,” *The Journal of Special Education*, vol. 41, no. 2, pp. 85–120, 2007.

[27] J. Hasbrouck and G. Tindal, “Oral reading fluency norms: A valuable assessment tool for reading teachers,” *The Reading Teacher*, vol. 59, no. 7, pp. 636–644, 2006.

[28] F. Claus, H. G. Rosales, R. Petrick, H.-u. Hain, and R. Hoffman, “A Survey about ASR for Children,” in *Proceedings of SLaTE*, Grenoble, 2013, pp. 26–30.

[29] N. J. Veenendaal, M. A. Groen, and L. Verhoeven, “What oral text reading fluency can reveal about reading comprehension,” *Journal of Research in Reading*, vol. 38, no. 3, pp. 213–225, 2015.

[30] M. C. Danne, J. R. Campbell, W. S. Grigg, M. J. Goodman, and A. Oranje, “Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading. The Nation’s Report Card. NCES 2006-469,” *National Center for Education Statistics*, 2005.