INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

ASR-based Evaluation and Feedback for Individualized Reading Practice

Yu Bai [1], Ferdy Hubers [1,2], Catia Cucchiarini [1], Helmer Strik [1,2,3]

[1] Centre for Language and Speech Technology (CLST), Radboud University Nijmegen, The Netherlands
[2] Centre for Language Studies (CLS), Radboud University Nijmegen, The Netherlands
[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands
y.bai@let.ru.nl, f.hubers@let.ru.nl, c.cucchiarini@let.ru.nl, w.strik@let.ru.nl

Abstract

Learning to read is a prerequisite to participate in our knowledge society. Developing reading skills requires intensive practice with individual evaluation and guidance by teachers, which is not always feasible in traditional classroom instruction. Automatic Speech Recognition (ASR) technology could offer a solution, but so far it has been mostly used to follow children while reading and to provide correct word forms through text-to-speech technology. However, ASR could possibly be employed at earlier stages of learning to read when children are still in the process of developing decoding skills. Early evaluation through ASR and individualized feedback could help achieve more personalized and possibly more effective guidance, thus preventing reading problems and improving the process of reading development.

In this paper we report on an explorative study in which an ASR-based system equipped with logging capabilities was developed and employed to evaluate decoding skills in Dutch first graders reading aloud, and to provide them with detailed, individualized feedback. The results indicate that ASR-based feedback leads to improved reading accuracy and speed and that the log-files provide useful information to enhance practice and feedback, thus paving the way for more personalized, technology-enriched approaches to reading instruction.

Index Terms: reading tutor, child speech, ASR, log-files, accuracy, fluency, speed, individualized feedback.

1. Introduction

Learning to read is one of the most fundamental skills children acquire at school. At the moment this is a topical issue as research indicates that many pupils experience difficulties in learning to read [1] and increasing numbers of pupils are functionally illiterate [2]. Learning to read requires intensive practice in reading aloud with individual guidance and feedback by teachers [3], which is difficult to realize in a classroom context. Educational software that incorporates Automatic Speech Recognition (ASR) technology has been proposed as an alternative as this can in principle provide automatic feedback on reading aloud, enabling pupils to practice more intensively, wherever and whenever they want [4]. Such software has been developed for English reading and turned out to be successful when tested in schools in the US and Canada [5], [6]. So far ASR technology has been mostly used to follow children while reading and to detect possible disfluencies so that support could be provided through text-to-speech technology to indicate the correct form of the word. However, ASR could possibly be employed at earlier stages of learning to read, when children are still in the process of developing decoding skills. This of course requires specialized algorithms that can detect reading errors at a more detailed level.

In this paper we report on an explorative study in which ASR technology was developed and employed to gain more insight into the possible contribution of ASR to automatized evaluation and feedback on reading aloud by Dutch children when they are in the process of acquiring decoding skills (grade 1). The ultimate aim of this research would be to develop more focused interventions and help realize more personalized reading practice.

2. Research Background

Although ASR technology is now employed in many devices and the appearance of applications like Apple Siri, Amazon Echo, Microsoft Cortana, and Google Home/Assistant might have suggested that the 'ASR problem' is solved, this does not mean that ASR can be easily employed in educational contexts for children, in particular in reading support. To evaluate and support reading skills ASR has to do more than simply recognizing the words a child is trying to say: It has to identify and diagnose errors and provide individualized feedback.

For English, there is a long history of developing ASR for reading assessment and instruction [5]–[8]. For Dutch, research on this topic [9]–[12] has been limited. Ideally, an Automatic Reading Tutor, should be able to monitor children while they read aloud and help when they encounter difficulties [6]. This requires more than simply recognizing the words: for each word read out by a pupil the algorithm should indicate whether it contains errors and where. This is a challenging task because this decision has to be made instantaneously, based on one single observation of the word produced (and e.g. not by aggregating scores on several words or instances of the same word), and because even human raters often disagree on what should be considered a mistake [13]. To the best of our knowledge, there are no systems that provide this kind of practice and automatic feedback at such detailed level.

3. Method

We developed reading software employing ASR technology that stores audio files (the utterances read by the children) and log files (containing large amounts of interesting information, e.g. the interactions of the children with the system) for further analysis. The software was tested with first graders in six Dutch primary schools. In the remainder of this section we first introduce the reading material employed in this study (3.1).

http://dx.doi.org/10.21437/Interspeech.2020-2842

Then we describe how the ASR backend interacts with the frontend (3.2) and how the user interface and feedback were designed (3.3). The experiments (participants and data analysis) are described in Section 3.4.

3.1. Reading Material

In many primary schools in the Netherlands children learning to read are first exposed to a considerable amount of explicit phonics instruction which is aimed at teaching them the basic structure of written language by showing the relationship between graphemes and phonemes [14]. The reading program Veilig Leren Lezen [15] is often used, in which children learn to read texts of increasing difficulty levels, with respect to text structure, vocabulary and length of words and sentences. In line with this practice, we selected material from the reading method for first graders by Zwijsen Publishers, Veilig & Vlot, which is used in the majority of primary schools in the Netherlands.

3.2. ASR Technology

The software consists of a front-end with the user-interface, and a back-end with the ASR engine. For the backend, we make use of the NovoLearning ASR. On the front-end words are shown on the screen. When the child clicks the microphone, the recording starts and the audio files are sent to the backend. The ASR then analyses the recorded spoken utterances, calculates scores (probabilities) at the phone and word level, which are expressed in numbers ranging from 0 to 100. The score at the word level is the minimum score of all the phones the word consists of. The scores are then used to provide feedback. If the score at the word level is lower than the threshold, the child gets feedback that the word was read incorrectly (see Figure 1). The threshold is now set to 50, but can be changed, even at the level of the individual child. All scores are stored in log-files, together with other relevant information such as the onset and offset of the speech, and the number of attempts. The audio and log-files are stored and are available for later analysis.

3.3. User Interface and Feedback

The software contains two types of exercises aiming at different reading skills: accuracy exercises and fluency exercises. The accuracy exercises focus on the pupils' reading accuracy of individual words and sentences. The pupil clicks the recording button and reads one word or sentence. With the ASR backend giving scores on each word, the software gives feedback on whether the target word or sentence is correct (see Figure 1). The pupil is asked to read the incorrect item again if the first attempt is incorrect. If the answer is correct, the recording button moves to the next item. If the second attempt is still incorrect, the software plays the correct word form and then asks the pupil to try again (3rd attempt). The pupil can listen to the correct speech by clicking the play button. If the answer is correct, the recording moves to the next item. If not, the feedback highlight remains on the current item and the recording button goes to the next item. The process is the same for sentences in the accuracy exercises. The feedback is shown on the incorrect words, but the pupil is asked to read the whole sentence again. In summary, after the 1st incorrect attempt, the pupil hears "try again"; after the 2nd attempt the correct form of the word or sentence is played; and after the 3rd incorrect attempt, only the incorrect word or sentence is highlighted.

The fluency exercises aim to improve the pupils' reading fluency while keeping track of accuracy at the same time. In the fluency exercise, pupils practice reading word lists and stories. In the word list exercises, fifteen words are shown on the screen. The pupil reads the word list in one go. Then feedback is shown to the pupil through a chart on which fluency is represented by the height of the rocket and accuracy is represented by a line of stars (see Figure 2).

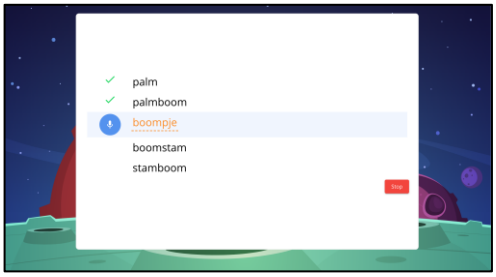

Figure 1: *Screenshot of an accuracy exercise.*

Then the screen goes back to the word list again and the correct forms of the incorrect words are played one by one. Subsequently, the pupil is asked to read the incorrect words by clicking the recording button one by one. Finally, the pupil reads the whole list again without receiving feedback. After that, the feedback chart is shown again with the feedback of both the first and the second attempt (see Figure 2). For stories, the difference from word list exercise is that after the first attempt, the pupil is asked to reread the sentences with one or more incorrect words.

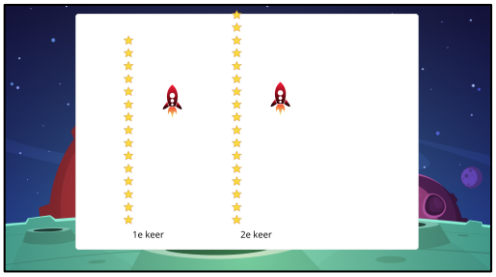

Figure 2: *Screenshot of a feedback chart.*

3.4. Experiments in Schools

3.4.1. Participants

We collected preliminary data from 39 Dutch pupils from Grade 1 in six primary schools. The pupils were between 6 and 7 years old and were in the early stages of learning to read.

3.4.2. Data analysis

In order to see whether the feedback was effective we ran linear mixed effects regression analyses (with pupils and words as random effects, intercept only) to analyse reading accuracy (operationalised as word probability) and reading speed (measured in graphemes per second) for the accuracy and fluency exercises. These analyses were conducted in the statistical software package 'R' version 3.4.0 [16], using the R packages 'lme4' [17], 'lmerTest' [18], and 'effects' [19].

4. Results

In this section, we present a general overview of the results (Section 4.1) with data of example pupils and the results of the reading accuracy and reading speed analyses (Section 4.2).

4.1. Overview

The 39 pupils who used our system read aloud 29007 words generating 5735 log files. We removed outliers, i.e. words with a reading speed (in graphemes/sec) higher than 25 as assessed by the ASR (4.18% of the data), because such a high reading speed probably implies that the user encountered problems when using the software. Table 1 shows the general information on the words from the two types of exercises.

Table 1: *General information on words in the exercises.*

Exercise Type	Word Count	Mean Probability	Mean Speed
Accuracy	10667	74.58 (SD=26.68)	7.62 (SD=3.73)
Fluency	17876	79.60 (SD=24.88)	8.22 (SD=3.88)

4.1.1. Problematic Words

We grouped together tokens of the same words that were read at the first attempt and calculated the mean probability scores of each word, which we define as reading accuracy. In Table 2 we present the ten most problematic words, i.e. with the lowest mean values, that were read at least 10 times in the log data.

Table 2: *10 most problematic words.*

Word	Count	Mean Probability	Standard deviation
stipje	10	40.39	32.98
strafpunt	10	42.76	31.64
strijken	11	43.68	34.92
broekrok	23	47.02	28.71
maud	20	47.94	25.56
druifje	20	48.06	29.94
welke	22	48.21	26.58
postzak	19	48.30	26.96
badmuts	12	48.45	24.07
tuinbank	10	49.80	37.91

4.1.2. Good Reader and Poor Reader

For each pupil we calculated the mean word probability at the first attempt. Pupils who read less than 100 words at the first attempt were filtered out.

Table 3: *10 most problematic words for Pupil 0112 (poor reader).*

Word	Count	Mean Probability	Standard deviation
bank	2	3.16	0.49
spons	2	3.90	4.54
klemt	3	12.49	0.28
zijn	2	19.06	17.88
nu	3	19.82	15.27
juich	7	21.85	29.15
reis	4	28.35	42.89
land	3	32.09	51.23
speelt	4	36.11	41.33
maar	2	36.99	42.59

We selected Pupil 0112, with a mean probability score of 40.12, as an example of a 'poor reader' and Pupil 0327, with a mean probability of 87.19, as an example of a 'good reader'. Table 3 shows ten problematic words for Pupil 0112. Table 4 shows ten problematic words for Pupil 0327.

Table 4: *10 most problematic words for Pupil 0327 (good reader).*

Word	Count	Mean Probability	Standard deviation
postzak	3	27.53	21.96
plank	3	39.90	38.07
schrik	2	44.74	58.21
postduif	3	47.57	30.79
plak	2	56.69	23.09
schrok	2	60.53	15.54
druifje	2	61.08	40.91
schik	2	63.79	39.58
je	3	64.14	27.98
plakstift	2	65.72	26.75

4.2. Reading accuracy and speed

4.2.1. Accuracy exercises

In these exercises we only included words that were read incorrectly twice by the same pupil, and thus had three attempts (n of words in isolation = 146, n of words in a sentence = 125). The regression analysis revealed that the pupils' reading accuracy significantly improved at the third attempt as compared to the first (*Beta* = 35.75, *SE* = 786.10, $p < .001$) and second attempt (*Beta* = 34.32, *SE* = 786.10, $p < .001$). See Figure 3 for a visualization of the effect.

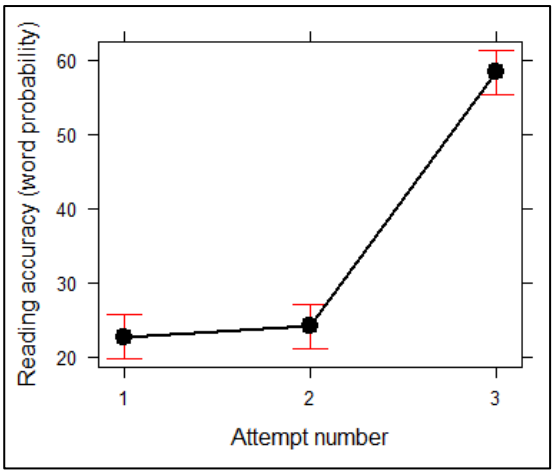

Figure 3: *The effect of the attempt number on reading accuracy in the accuracy exercises.*

With respect to reading speed we only found a significant effect of the context in which the word was presented. Words embedded in a sentence were read faster than words read in isolation (*Beta* = 1.75, *SE* = 0.39, $p < .001$). However, pupils did not improve in terms of reading speed at the third attempt as compared to the first (*Beta* = 0.31, *SE* = 0.41, $p = .441$) and second attempt (*Beta* = 0.32, *SE* = 0.40, $p = .434$).

4.2.2. *Fluency exercises*

In these exercises we only included words that were read twice by the same pupil (*n* of words in a word list = 2612, *n* of words in a story = 4005). The regression analysis showed a significant improvement of reading accuracy, which was larger for words in word lists (*Beta* = 3.39, *SE* = 0.46, $p < .001$) than for words in a story (*Beta* = 0.91, *SE* = 0.46, $p < .05$), as appears from Figure 4.

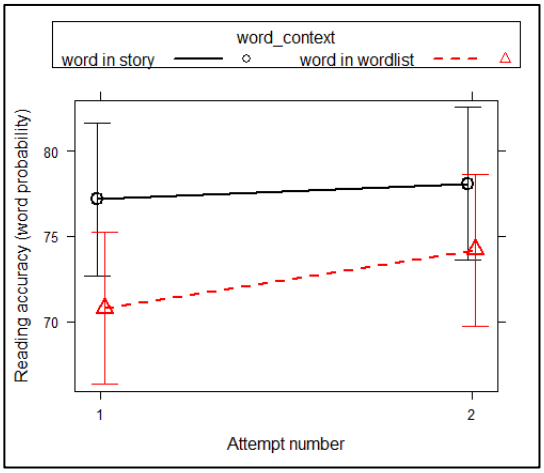

Figure 4: *The effect of the attempt number on reading accuracy in the fluency exercises.*

The analyses of reading speed revealed that pupils read the same words faster at the second attempt as compared to the first one (*Beta* = 0.66, *SE* = 0.01, $p < .001$) (see Figure 5). Moreover, a general effect of the context was found in which the word was presented. Words presented in a story were read faster than words presented in a word list (*Beta* = -2.46, *SE* = 0.01, $p < .001$).

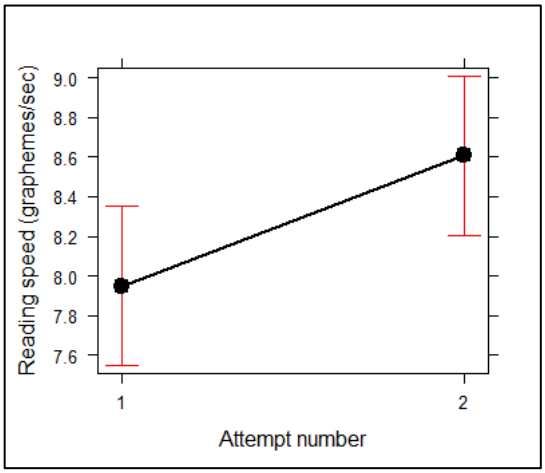

Figure 5: *The effect of the attempt number on reading speed in the fluency exercises.*

5. Discussion and Conclusions

Learning to read is one of the core tasks of primary education, because well-developed reading skills are crucial for the school career and for participating in our knowledge-based society [20]. Reading literacy has recently gained additional attention because of concerns about the current levels of reading literacy [13], [21] and functional literacy [1] among Dutch pupils.

In this paper we have proposed a new ASR-based approach for early reading practice that allows to identify reading difficulties at an early stage and to provide individualized feedback. The proposed method is innovative as digitalization in Dutch reading instruction has so far been limited to drag and drop exercises [14] while previous research on ASR for Dutch child speech did not provide speech diagnostics and tailored feedback [7]–[9], [15].

The results of this exploratory study show that through this novel approach it is possible to identify reading errors in the early stages of the process, to track pupils' performance and to collect log data for analysis. Through analyses of the log data we can study how feedback can contribute to improving progress in terms of reading accuracy and speed, which words are problematic for most pupils, which words are problematic for individual pupils, and which pupils are good or poor readers. An option is to make the system more tolerant for poor readers, by lowering the threshold. This is an interesting option, as too much feedback on errors could demotivate pupils.

In the accuracy exercises we saw that reading accuracy significantly improved at the third attempt after feedback had been provided that highlighted the incorrect word and that played a recording of the correct form of the erroneous word. We observed that words embedded in a sentence were read faster than words in isolation, and that reading speed did not improve at the third attempt as compared to the first and second attempt. The latter is not surprising, as the goal of these accuracy exercises is to improve accuracy, and not fluency.

In the fluency exercises we did observe that speed improved after feedback: pupils read the same words faster at the second attempt than at the first attempt. Words presented in a story were read faster than words presented in a word list. We also observed a significant improvement of reading accuracy, which was larger for words in word lists than for words in a story.

These results are interesting for the development of educational applications focusing on reading development, but also for educational research on the reading process. Although the present study was limited as it included a relatively small number of pupils, its results are promising. We are now collecting more data. In the near future this approach could be applied to develop educational apps for young primary school children in which they read words and text aloud, their speech is recorded and analyzed through ASR to automatically provide information about their reading errors so that suitable remedial exercises can be suggested.

6. Acknowledgements

We would like the thank our colleagues Marjoke Bakker and Erik van Schooten as well as our partners in the DART project [http://hstrik.ruhosting.nl/DART/]: NovoLearning [https://www.novo-learning.com/], esp. Joost van Doremalen and David van Leeuwen; and Zwijsen publishers [https://www.zwijsen.nl/], esp. Rosemarie Irausquin and Martin de Jong. Special thanks go to all the children who participated, their parents and their teachers. This work (project number 40.5.18540.121) is funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), the Dutch Organization for Scientific Research.

7. References

[1] I. V. S. Mullis, M. O. Martin, P. Foy, and M. Hooper, "PIRLS 2016 International Results in Reading," 2017.

[2] R. C. W. Feskens, H. Kuhlemeier, and G. Limpens, "Resultaten PISA-2015 in vogelvlucht. Praktische kennis en vaardigheden van 15-jarigen," Arnhem, 2016.

[3] National Reading Panel, "Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction," 2000.

[4] M. J. Adams, "The promise of automatic speech recognition for fostering literacy growth in children and adults," in *Handbook of literacy and technology, Volume 2*, M. McKenna, L. Labbo, R. Kieffer, and D. Reinking, Eds. Hillsdale, NJ: Lawrence Erlbaum, 2005.

[5] J. Mostow, J. Nelson-Taylor, and J. E. Beck, "Computer-Guided Oral Reading versus Independent Practice: Comparison of Sustained Silent Reading to an Automated Reading Tutor That Listens," *J. Educ. Comput. Res.*, vol. 49, no. 2, pp. 249–276, Sep. 2013.

[6] K. Reeder, J. Shapiro, J. Wakefield, and R. D'Silva, "Speech Recognition Software Contributes to Reading Development for Young Learners of English," *Int. J. Comput. Lang. Learn. Teach.*, vol. 5, no. 3, pp. 60–74, Aug. 2015.

[7] B. Wise, R. Cole, S. Van Vuuren, S. Schwartz, L. Snyder, N. Ngampatipatpong, and J. Tuantranont, "Learning to Read with a Virtual Tutor: Foundations to Literacy," in *Interactive Literacy Education: Facilitating literacy learning environments through technology*, C. Kinzer and L. Verhoeven, Eds. Mahwah, NJ: Lawrence Erlbaum, 2005.

[8] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: The role of multiple information sources," in *2007 IEEE 9Th International Workshop on Multimedia Signal Processing, MMSP 2007 - Proceedings*, 2007, pp. 26–30.

[9] L. Cleuren, "Elements of Speech Technology Based Reading Assessment and Intervention," 2009.

[10] J. Duchateau, L. Cleuren, H. Van Hamme, and P. Ghesquière, "Automatic assessment of children's reading level," in *Interspeech 2007*, 2007, pp. 1210–1213.

[11] J. Duchateau, Y. O. Kong, L. Cleuren, L. Latacz, J. Roelens, A. Samir, K. Demuynck, P. Ghesquière, W. Verhelst, and H. Van Hamme, "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules," *Speech Commun.*, vol. 51, no. 10, pp. 985–994, Oct. 2009.

[12] M. Nicolao, M. Sanders, and T. Hain, "Improved Acoustic Modelling For Automatic Literacy Assessment Of Children," in *Interspeech 2018*, 2018.

[13] P. Price, J. Tepperman, M. Iseli, T. Duong, M. Black, S. Wang, C. K. Boscardin, M. Heritage, P. David Pearson, S. Narayanan, and A. Alwan, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Commun.*, vol. 51, no. 10, pp. 968–984, Oct. 2009.

[14] H. Wentink, "From graphemes to syllables," University of Nijmegen, 1997.

[15] M. J. C. Mommers, L. Verhoeven, and S. Van der Linden, *Veilig Leren Lezen*. Tilburg: Zwijsen, 1990.

[16] R Development Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2008.

[17] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting Linear Mixed-Effects Models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[18] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J. Stat. Softw.*, vol. 82, no. 13, 2017.

[19] J. Fox, "Effect Displays in R for Generalised Linear Models," *J. Stat. Softw.*, vol. 8, no. 15, pp. 1–27, 2003.

[20] K. Berkling and U. Reichel, "Progression in Materials for Learning to Read and Write - a CrossLanguage and Cross-Century Comparison of Readers," in *WOCCI-2016*, 2016, pp. 1–9.

[21] Radboud Recharge, "Virtual speech coach helps children learn to read," 2019. [Online]. Available: https://www.radboudrecharge.nl/en/article/virtual-speech-coach-helps-children-learn-to-read.