An ASR-based Reading Tutor for Practicing Reading Skills in the First Grade: Improving Performance through Threshold Adjustment

Yu Bai[1], *Ferdy Hubers*[1,2], *Catia Cucchiarini*[1], *Helmer Strik*[1,2,3,4]

[1] Centre for Language and Speech Technology (CLST), Radboud University Nijmegen
[2] Centre for Language Studies (CLS), Radboud University Nijmegen
[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen
[4] NovoLearning B.V., Nijmegen

y.bai@let.ru.nl, f.hubers@let.ru.nl, c.cucchiarini@let.ru.nl, w.strik@let.ru.nl

Abstract

Automatic Speech Recognition (ASR) technology can potentially be employed to provide intensive practice and feedback to young children learning to read. So far there has been limited research on the use of ASR in the early stages of learning to read when children are still developing decoding skills. For this purpose, we developed an ASR-based system equipped with logging capabilities that can evaluate decoding skills in Dutch first graders reading aloud and provide them with instantaneous feedback. In a previous study we found that ASR-based feedback led to improved reading accuracy and speed, and that useful information could be obtained from the log-files. For the present paper we conducted thorough analyses of the data obtained with this ASR-based system by comparing it to human annotations of the same read aloud 11849 words from 38 pupils. We present the results of our analyses, and discuss how they can contribute to better and more personalized ASR-based reading instruction.

Index Terms: Automatic Speech Recognition, reading tutor, child speech

1. Introduction

Automatic Speech Recognition (ASR) technology has previously been incorporated in education software for learning to read, because of its potential to provide intensive practice in reading aloud and, possibly, to detect reading problems. As a consequence, ASR technology has been mostly used to follow children while they read aloud a text and to identify upcoming reading difficulties, for instance because children hesitate to read specific words. In turn, support could be provided to teach pupils the correct form by resorting to text-to-speech technology. In our own research we decided to investigate the usability of ASR at earlier stages of learning to read, when children are still acquiring decoding skills. Within the framework of the DART project (https://dart.ruhosting.nl/), we developed an ASR-based system equipped with logging capabilities that can evaluate whether Dutch first graders reading aloud read words and sentences correctly. In turn the system provides feedback and opportunities for rehearsal so that pupils can practice and improve both reading accuracy and speed. In a previous study [1] we found that the feedback provided by our system managed to improve both reading accuracy and speed. In addition, the log-files provided useful insights that could be employed to improve both practice and feedback. One aspect that was not investigated in that study was the performance of the ASR-based system in establishing whether words were read correctly or incorrectly. In the present paper we address this issue by comparing the performance of our system to annotations of the same read aloud speech obtained from 38 pupils. The aim is to gain insights into the performance of our system with a view to improving it, for instance by applying different thresholds at which a word is judged to be incorrect. In addition, these more detailed analyses can provide more specific information on which letters and sounds are more problematic for children to read and for the ASR to recognize, respectively. Both types of insights can increase our understanding of the nature of reading errors and ASR errors and can thus contribute to developing better and more personalized ASR-based reading instruction.

2. Research Background

Several studies have investigated the contribution of ASR technology in supporting children learning to read. Most of the research addressed English reading skills [2]–[6], while studies on other languages and Dutch in particular have been limited [7]–[11]. Most of the commercial applications that have been developed so far aim at monitoring children while they read aloud and at providing support when they hesitate, for example because they do not know how a word should be read aloud. For English, commercial products exist that employ ASR in this way like the Reading Assistant (http://www.readingassistant.com/), IBM Reading Companion (https://www.ibm.com/ibm/responsibility/downloads/initiatives/ReadingCompanion.pdf), and the ReadingBuddy (http://readingbuddysoftware.com/#).

However, we think that ASR could provide a valuable contribution in earlier stages of learning to read, when children are developing decoding skills [12], that is when they learn how to convert letters to speech sounds, and need to focus on both accuracy and fluency [13]. In these stages ASR can be employed to determine whether children manage to read words and sentences correctly [2]–[4], but this requires dedicated algorithms that can identify reading errors at more detailed levels. In addition, if feedback has to be provided instantaneously, then the decision whether the word was read correctly or not also has to be instantaneous and it has to be based on single observations, which, of course, is quite challenging. To develop this kind of system it is extremely important to gain insight into ASR performance to determine to what extent these decisions are made appropriately by the system. To get an idea of the challenging nature of making such decisions, it is also important to note that even human raters often disagree on what should be considered a reading error

10.21437/IberSPEECH.2021-3

[14]. So far there has been limited research that has compared ASR performance to human performance at this level of detail in an online system, which is understandable in view of the time-consuming nature of the annotations required. The methodology applied in the current study, which is partly similar to the one described in [1], is presented below.

3. Method

3.1. Reading tutor

The reading tutor was developed to follow as much as possible the reading method for first graders developed by Zwijsen Publishers, 'Veilig Leren Lezen', which is the most widely used reading method in Dutch primary schools [15]. Two types of exercises were incorporated which address 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 on the recording button and reads one word or sentence. With the ASR backend giving scores on each word, the reading tutor gives feedback on whether the target word or sentence is correct. If the target word or sentence is read incorrectly, the pupil has to try again (up to a maximum of three attempts).

The fluency exercises aim to improve the pupils' reading fluency while keeping track of accuracy at the same time. In the fluency exercises, pupils practice reading word lists and stories. They are instructed to read a word list or a story in one go (the first try). Subsequently, they receive feedback and are prompted to try the incorrect words or sentences again. Next, pupils have to read the same word list or story again (the second try) and are encouraged to read faster with the same accuracy.

See [1] for a detailed description of the system and the reading materials used in the system.

3.2. ASR Technology

The reading tutor uses the NovoLearning [https://www.novo-learning.com/] ASR engine, which analyses the spoken attempts by the children by calculating scores (probabilities) at the phone and word level. These scores 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, and is used to provide feedback. If this score is lower than the threshold, the child gets feedback that the word was read incorrectly. For the current data collection, the threshold was set to 50. 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.

3.3. Data Collection

38 Dutch pupils from Grade 1 in six primary schools together read 28543 words. The pupils were between 6 and 7 years old and were in the early stages of learning to read. The experiment was conducted during the COVID-19 pandemic, so most pupils used the reading tutor at home, in an uncontrolled context.

3.4. Manual Transcription and DP alignment

Out of the 28543 words, 12185 words (42.69%) were orthographically transcribed by a human annotator who was familiar with the procedure adopted in Dutch primary schools to score tests of reading accuracy and fluency. From these words, 4095 words were taken from the accuracy exercises and 8090 words from the fluency exercises. From the accuracy exercises, we selected words with a third attempt (meaning that the first two attempts by the same pupil were judged to be incorrect, as explained in Section 3.1). The first and second attempt of that specific word were also included. To balance the number of correct and incorrect words judged by the software, the same words that were judged as correct at the first and second attempts were also selected for transcription. From the fluency exercises, word lists and stories at the first and second try were selected. To balance the number of correct and incorrect words, the same words that were judged as correct in the fluency exercises were selected as well. The transcriber made orthographic transcriptions of words coming from all four types of exercises. The procedure was as follows: First, Praat [16] textgrids containing the prompt were automatically generated using a Python script. Next, the transcriber used Praat to verify whether the prompt had actually been read. If this was the case, she did not have to change anything and could move on to the next textgrid. If the prompt was not a correct orthographic transcription of what had been read, the transcriber was instructed to change the orthographic transcription in line with what she had heard.

Figure 1: *The procedure of DP alignment.*

After the manual transcription had been obtained, we used the dynamic programming algorithm ADAPT [17] to align the manual transcription with the prompt. Figure 1 presents an example of how the prompt and manual transcription were aligned and how the words were judged to be correct based on the manual transcription. It is important to note that this process may not be straightforward in certain cases. For instance, a child may read the words a couple of times and sometimes the correct word is produced only at the end of the utterance (see Figure 1). In such cases it is important that the ASR detects the correct word in the whole utterance, because this is the way teachers normally judge such reading attempts. In establishing the degree of correspondence between the manual transcription and the prompt this should be also taken into account. For example, if the reader tried to read a word multiple times, the last attempt should be aligned to the word in the prompt. To achieve this, we used the ADAPT algorithm in inverse direction and aligned the prompt and the manual transcription from right to left instead of from left to right (STEP 1). In this way, the words in the prompt were aligned with the last occurrence of the word in the manual transcription, which means that the last trial of a word in the prompt was recognized. Words in green in Figure 1 are pronounced correctly in the last trial. For the words that

were not aligned (words in red in Figure 1), we used the normal ADAPT without reversion so that more words were aligned (STEP 2). The 'FINAL RESULT' is a binary score for each word: 0 – incorrect, 1 – correct. [Bo is a boy's name.]

3.5. Data Analysis

To gain insight in the performance of our system, we compared the system-based judgements with the human-based judgements of word reading by applying different thresholds. To this end, we calculated different measures for thresholds varying from 0 to 100. Cohen's kappa was calculated as a measure of agreement between the system-based and the human-based judgements [18], and we also calculated measures such as the percentages of correct acceptance (CA), correct rejection (CR), false acceptance (FA) and false rejection (FR), precision (P), recall (R), and the F-measure as in [19] to get insight into the quality of the feedback. This terminology and related measures were preferred because they are more transparent than those indicating false positives, negatives, etc., which can vary depending on the perspective from which the binary classification is analyzed [20].

4. Results

In this section, we present a general overview of the results concerning the ASR scoring (Section 4.1) and then go on to investigate how optimal thresholds can be established to improve the performance of the ASR system (Section 4.2).

4.1. General results of ASR scoring

12185 words read by 38 pupils were transcribed. For 2.8% of these words, no transcription could be provided, because the corresponding audio recordings were empty. We excluded these words so that in total 11849 words remained for the analysis. Figure 2 shows the frequency count of the word probability scores given by the ASR backend in the system. The majority of words received a high word probability, meaning that it is highly likely that these words were read correctly.

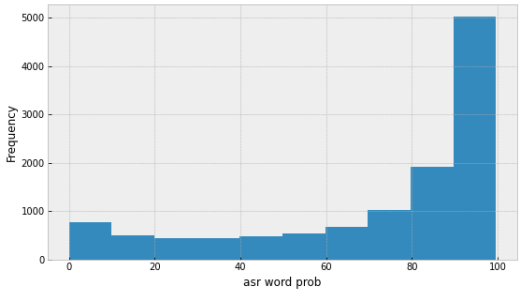

Figure 2: *Distribution of the word probabilities.*

4.2. Optimal ASR thresholds

Figure 3 shows Cohen's kappa as a function of the threshold ranging from 0 to 100. The optimal threshold is 48. Using this threshold gives the highest agreement between the system-based judgements and the human-based judgements. Cohen's kappa at a threshold of 48 is .41, which indicates moderate agreement [21].

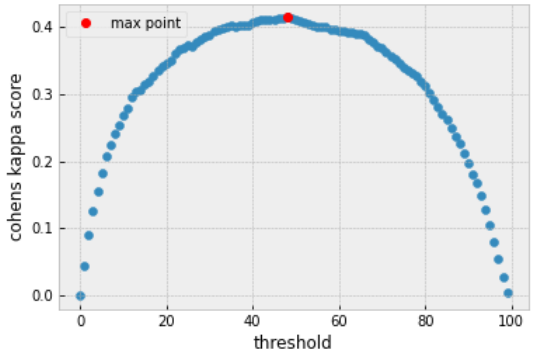

Figure 3: *Cohen's kappa for different thresholds.*

Based on the word probability scores, the reading tutor classifies words into correct and incorrect with a threshold that is manually set from 0 to 100. This classification results in four outcomes: correct acceptance (CA), correct rejection (CR), false acceptance (FA) and false rejection (FR). Figure 4 shows the percentages of CA, CR, FA and FR for thresholds from 0 to 100. At a threshold of 36, the percentages of false rejects, correct rejects and false accepts are about equal. The percentage of correct accepts is very high in general, but seems to drop rapidly after a threshold of about 70.

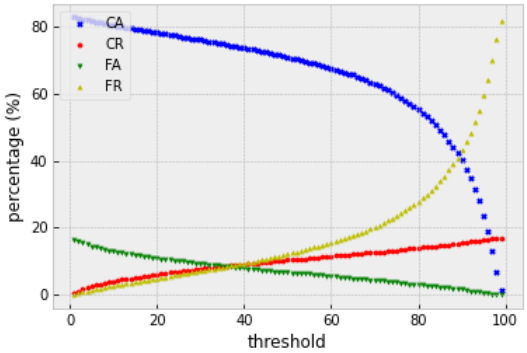

Figure 4: *CA, CR, FA and FR for different thresholds.*

Figure 5 shows F-measures of CA and CR for thresholds from 0 to 100. The threshold with the highest F-measure of CR (52.51%) is 48, which is close to the optimal threshold based on Cohen's kappa and which leads to an F-measure of CA that is still above 80%.

Table 1a shows the percentages of CA, CR, FA and FR for different thresholds from low to high. We included two extreme thresholds (20 and 60), a threshold of 36 at which the error rate (FA and FR) is equal, the optimal threshold based on Cohen's kappa and the F-measure for CA (48), and the threshold that we used in our data collection (50).

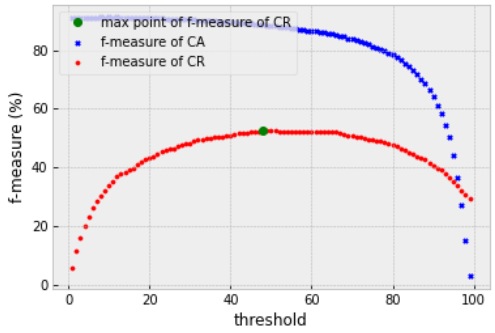

Figure 5: *F-measures of CA and CR as a function of different thresholds.*

As can be expected, we see in this table that when the applied threshold is stricter the percentage of CA decreases, while the percentage of CR increases. In addition, the percentages of FA and FR are generally low. Even with a threshold of 60, the percentage of false rejects is far below 20%.

We calculated the mean probability scores for each pupil who read over 300 words. An example of a poor reader with a mean probability of 54.796% and an example of a good reader with a mean probability of 84.257% were selected. The maximum Cohen's kappa (0.349) for the poor reader is at a threshold of 35, while the maximum Cohen's kappa (0.371) for the good reader is at a threshold of 56.

Table 1: *CA, CR, FA and FR at different thresholds, (a) average for all readers, (b) poor, (c) good reader*

Readers	**Thresholds**	**CA (%)**	**CR (%)**	**FA (%)**	**FR (%)**
		(a)			
All readers	20	78.4	6.0	10.9	4.8
	36	74,6	8.5	8.4	8.4
	48	71.5	10.1	6.8	11.6
	50	70.9	10.3	6.6	12.2
	60	67.4	11.4	5.5	15.6
		(b)			
Poor reader	20	62.7	11.5	15.3	10.5
	36	57.0	15.7	11.1	16.2
	48	51.1	17.7	9.1	22.1
	50	49.4	17.9	8.9	23.8
	60	40.6	19.1	7.8	32.5
		(c)			
Good reader	20	88.5	1.6	8.0	2.0
	36	87.4	2.7	6.9	3.0
	48	85.7	3.6	6.0	4.7
	50	85.6	3.7	5.9	4.8
	60	83.9	4.4	5.2	6.6

Table 1b and Table 1c show the percentages CA, CR, FA and FR of the poor reader and the good reader at different thresholds. When comparing these values, it can be seen that changing the threshold especially leads to substantial changes in all measures for a poor reader. For example, raising the threshold from 20 to 60, increases the percentage of FR by 22%. For a good reader, however, the measures do not change much as the threshold increases. Here changing the threshold from 20 to 60 only results in a 4.6% increase of FR.

5. Discussion and Conclusions

Within the DART project, we developed an ASR-based reading tutor that provides instantaneous feedback on the correctness of words and sentences read aloud by first graders, to be used in carefully controlled experiments in schools. Because of COVID 19 these experiments could not be conducted and children were allowed to use the system at home, in an uncontrolled context. An important requirement of the reading tutor was that it should ignore initial, often incomplete, attempts at reading the words, and should evaluate only the last attempt. The results show that the reading tutor was capable of doing this. However, there is room for improvement, for instance, by improving the acoustic models and language model of the ASR, and by studying how to obtain an optimal score for the whole word (now we use the minimum probability of all the phones composing the word).

In the current paper we investigated the effect of applying different thresholds on the performance of this online reading tutor. The results show that for all our data together, the maximum value of Cohen's Kappa and F(CR) are at a threshold of 48, which is very close to the threshold of 50 which was determined in the pilot phase and eventually adopted in the experiment proper. Increasing the threshold makes the system stricter and leads to higher values of CR and FR and lower CA and FA. Between 40 and 60, the changes are small and above 70 large changes are observed. In terms of feedback errors, the balance between FA and FR is important, as there is a trade-off between these two types of errors. Missing errors (FA) leads to reduced corrective feedback and possibly to less effective feedback. Flagging correctly read words as being incorrect (FR) can cause frustration and demotivation. In our informal observations during the pilot experiments, we noticed that the latter was indeed the case. So a good strategy in choosing the threshold probably is to prioritize reducing FR. The performance obtained in our experiment is comparable to that achieved in previous research on Dutch ASR-based reading assessment [11], with the important difference that that was a corpus-based study, while the present research refers to an online system used in an uncontrolled environment. Deciding what the optimal threshold should be in practice, will depend also on many other factors, such as the proficiency level of the child (see Table 1), the difficulty level of the content, the goals of the system, and preferences of the teachers and the pupils. An important final observation is that although the performance of the current system can still be enhanced, this was in any case sufficient to bring about improvement, as we saw in [1], and children were already positive about the current system.

6. Acknowledgements

We are grateful to Anna Krispin for transcribing the data and to Wieke Harmsen for helping us with the ADAPT alignment and the automated scoring. In addition, we would like the thank our colleagues Marjoke Bakker and Erik van Schooten as well as our partners in the DART project [http://dart.ruhosting.nl/]: 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] Y. Bai, F. Hubers, C. Cucchiarini, and H. Strik, "ASR-Based Evaluation and Feedback for Individualized Reading Practice," in *Interspeech 2020*, 2020, pp. 3870–3874.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 1015–1028, 2011.

[7] L. Cleuren, "Elements of Speech Technology Based Reading Assessment and Intervention," Ph.D dissertation, KU Leuven, Belgium, 2009.

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

[9] 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.

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

[11] E. Yilmaz, J. Pelemans, and H. Van Hamme, "Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model," in *Interspeech 2014*, 2014, pp. 969–972.

[12] A. Castles, K. Rastle, and K. Nation, "Ending the Reading Wars: Reading Acquisition From Novice to Expert," *Psychol. Sci. Public Interes.*, vol. 19, no. 1, pp. 5–51, 2018.

[13] J. J. Pikulski and D. J. Chard, "Fluency: Bridge Between Decoding and Reading Comprehension," *Read. Teach.*, vol. 58, no. 6, pp. 510–519, 2005.

[14] 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.

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

[16] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 30 December 2020 from http://www.praat.org/.

[17] B. Elffers, C. van Bael, and H. Strik, "ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions," Internal report, University of Nijmegen, 2005.

[18] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[19] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study.," in *Speech & Language Technology in Education -SLaTE*, 2009, no. 2, pp. 2–5.

[20] D. M. W. Powers, "Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation," Adelaide, Australia, 2007.

[21] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, 1977.