EVALUATING THE ACCURACY OF DECODING IN CHILDREN WHO READ ALOUD

E. Bruno[1], S. Giulivi[2], C. Cappa[3], M. Marini[4], M. Ferro[1]

[1]Institute for Computational Linguistics ILC-CNR Pisa, Italy

[2]University of Applied Sciences and Arts of Southern Switzerland SUPSI Locarno, Switzerland

[3]Institute for Clinical Physiology IFC-CNR Pisa, Italy

[4]Department of Information Engineering, University of Pisa, Pisa, Italy

ester.bruno@ilc.cnr.it, sara.giulivi@supsi.ch, cludia.cappa@cnr.it,
marco.marini@phd.unipi.it, marcello.ferro@ilc.cnr.it

Abstract: **Digital tools based on automatic speech recognition (ASR) could be a useful support for teachers in assessing the reading skills of the students. We focus on the evaluation of the decoding accuracy of children with grade level ranging from the 3$^{\text{rd}}$ to the 6$^{\text{th}}$ performing a reading aloud task on a narrative text displayed on an ordinary tablet using the ReadLet platform. On the basis of previously collected data, we built a gold dataset with sentences characterised by the audio data, the original text to be read, and the text actually spoken by the child. By using the open-source Kaldi toolkit an ASR system based on the GMM-HMM model was trained on the training portion of the gold dataset. The accuracy of the ASR system was calculated as the ability to correctly decode the test audio data with respect to the annotated text, and the decoding accuracy of the children was estimated by measuring the gap between the results obtained with the annotated text and the original text. A consistent trend with increasing grade level was found in terms of word correctness, substitutions and insertions, while the trained model appears to be significantly able to evaluate the children decoding accuracy.**

Keywords: **speech recognition, decoding accuracy, reading aloud, voice parameters, Kaldi, GMM-HMM acoustic model**

I. Introduction

Reading and understanding a written text are among the most relevant skills in everyone's life [1]. Whether it is to study, to read for personal pleasure, to obtain information, to use instructions, to find communications or updates, we are faced with the need to access the content of a written text. The results of the OECD-PISA 2018 international survey is the most recent in which reading skills were the main area of investigation, and return an uncomfortable international picture, from which Italy does not differ [2]. The assessment of reading skills is achievable by the educational institutions, and the combination of NLP and ICT technology can substantially help the teachers in this task [3].

The process of decoding and understanding during reading were considered by the American Psychiatric Association 2013 as two independent processes, however able to influence each other [4]. The assessment of such processes in ecological conditions on primary school children is the objective of the AEREST protocol [5], which is implemented into the ReadLet platform [6] so that, by using an ordinary tablet, the reading efficiency is automatically evaluated as the integration of the ability to decode and understand a text.

II. Materials and methods

The AEREST protocol provides for the administration of narrative-descriptive texts in three decoding modalities: silent reading, reading aloud, and listening. The decoding step is followed by a questionnaire to evaluate the comprehension of the text just read. By using an ordinary tablet, ReadLet takes care of recording the speech produced by the child, keeps track of child's finger movement on the screen and, finally, stores the answers given to the comprehension questionnaire. All acquired data are aligned over time. Three contributions are calculated to evaluate the reading efficiency of the child:

i) the decoding speed, ii) the correctness of the reading and iii) the understanding of the text. Points i) and iii) are already fully automated within the ReadLet platform and in this article we focus on point ii), with the aim of creating a tool that is able to automatically draw the decoding accuracy in terms of correct words, deletions, substitutions and self-corrections.

As part of the AEREST project in 2019, we created a gold dataset starting from the data acquired using from 419 children with a grade level between the third and the sixth. The overall database includes 419 reading-aloud trials and a total of 13118 sentences. To create the gold dataset, a first step involved the selection of the trials in which the child marked the text with the finger for at least 70% of the text length. Since the speech and the finger tracking data were simultaneously recorded during the trial and subsequently aligned over time, we relied on the finger tracking data to automatically split the audio data into sentences. The audio segmentation was then refined manually by means of an ad-hoc audio editing tool and, additionally, the annotation was augmented by taking into account the text actually spoken by the child compared to the original sentence.

From ReadLet we obtained a gold dataset composed by 873 sentences characterized as i) the audio data (i.e. the speech of the child), ii) the original sentence (i.e. the text that should have been pronounced by the child), and iii) the annotated sentence (i.e. the transcription of the actual speech of the child).

The ReadLet dataset was integrated with the CLIPS dataset, 16120 recordings about 8 hours and 30 minute from 250 adult subjects [7]. Once the total dataset was obtained, training and testing of an ASR system based on the GMM-HMM model [8] was performed using the open-source Kaldi toolkit [9]. The GMM-HMM model is composed by 15019 gaussians and it has been trained with the Speaker Adaptive Training (SAT) algorithm [10]. The feature vector was projected by Linear Discriminant Analysis criterion and transformed by Maximum Likelihood Linear Transformation [11] (LDA + MLLT + SAT). The final vector consisted of 40 features. MFCC features were extracted from the audio data and the decoding was performed on the fully expanded decoding graph (HCLG) that represents the language-model, pronunciation dictionary (lexicon), context-dependency, and HMM structure. Both mono-phone and tri-phone model were run and, since the latter outperformed the mono-phone model, we will focus on the tri-phone model only.

Finally the training set was obtained by all CLIPS recordings plus the 60% of the gold dataset, while the test set was built with the remaining 40% of the gold dataset. The random selection of the training and testing datasets was repeated 5 times and the results were averaged accordingly.

We trained the ASR system by feeding the model with the audio data and the annotated sentences belonging to the training dataset. During testing, we fed the model with the testing audio data and we compared the ASR transcriptions with two kind of references: i) the annotated sentences and ii) the original sentences.

III. Results

The predictions of the model run on the test audio data were compared to the target text. The accuracy of the ASR was first measured by Word Error Rate (WER) which is computed as the overall number of predicted words not matching the target text, divided by the number of total words. The preliminary results of the model show a mean WER equal to 10.95% (std=2.00%). Going more in deep, for each grade level the accuracy was evaluated as i) the average number of words per sentence correctly recognised by the model (correctness), ii) the average number of words per sentence substituted into the target text (substitutions), iii) the average number of words per sentence removed from the target text (deletions), and iv) the average number of words per sentence added to the target text (insertions). By using the annotated sentences as the target text we obtained the results shown in Fig. 1.

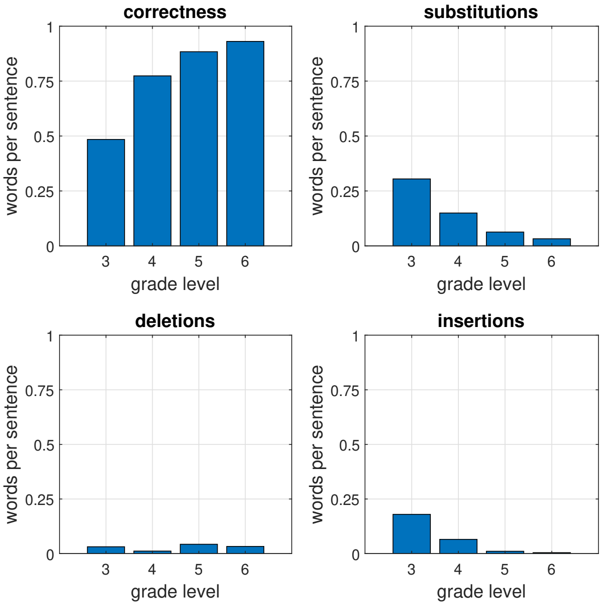

Figure 1: Accuracy of ASR system fed with the test audio data and using the annotated sentences as the target text. For each grade level, the average number of correct/substituted/deleted/inserted words per sentence is shown.

The model accuracy was also calculated on the same test audio data using the original sentences as the target text. The difference of the correctness obtained using the two target texts (i.e. the correctness on the annotated text minus the correctness on the original text) is shown in Fig. 2. While the correctness of the model on the annotated text should tell us about the accuracy of the ASR system itself, the difference of such correctness with the one obtained on the original sentences should tell us about the performance of the children.

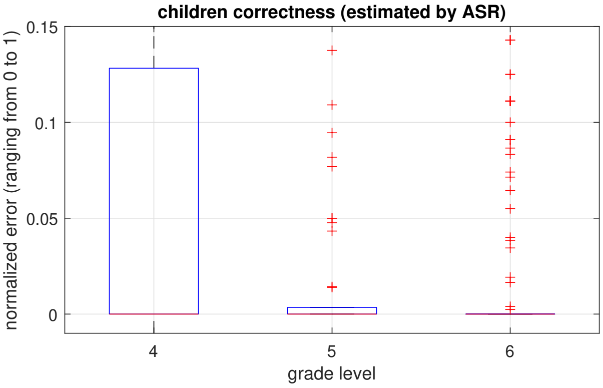

Figure 2: Children decoding accuracy estimated by the ASR system, expressed as the average number of misspelled words per sentence and calculated as the difference between the ASR correctness on the annotated sentences (Fig. 1 top-left) and the ASR correctness on the original sentences.

Finally, we evaluated the normalised edit distance between the annotated and the original sentences to obtain the reference correctness baseline for comparing the correctness estimated by the ASR system (see Fig. 3).

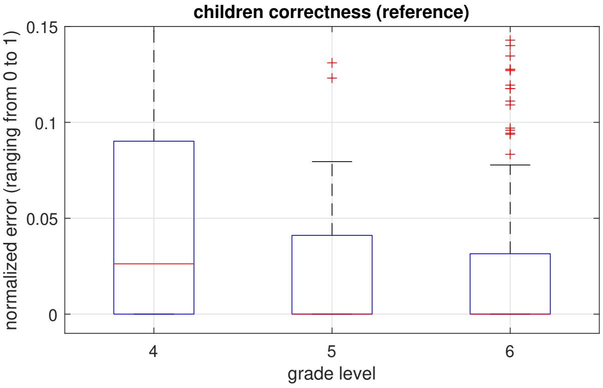

Figure 3: Children decoding accuracy used as reference, calculated as the normalised edit distance between the annotated sentences and the original sentences in the test set.

To validate the results provided by the ASR system about the performance of the children in correctly decoding the text during the reading aloud task, we calculated the Spearman rank correlation between the data shown in Fig. 2 and in Fig. 3. For each grade level, the correlation value along with the statistical significance is shown in Table 1.

grade level	r	p-value
4	0.63	$<10^{-3}$
5	0.50	$<10^{-5}$
6	0.66	$<10^{-10}$

Table 1: Spearman rank correlation between the children accuracy in decoding estimated by the ASR shown in Fig. 2 and the decoding accuracy calculated on the basis on the manually annotated sentences shown in Fig. 3. For each grade level the correlation value is shown together with its statistical significance.

IV. Discussion

As it can noticed in Fig. 1, the accuracy of the ASR model in terms of correctness is around 50% on 3rd graders, while the accuracy grows to 90% on 6th graders. The trend of substitution and insertion statistics goes in the same direction, showing that the more the reader is skilled, the more the model is able to predict the annotated text which, by definition, should reflect the audio data. Anyway a number of factors (e.g. the limited dataset, the poor annotation, the noisy audio, the poor fluency of the reader among all) may prevent the model to gain the 100% accuracy. For grade levels where the accuracy of the model is above 75% (i.e.. grade level ranging from 4 to 6) we show in Fig. 2 the evaluation of the accuracy of the children by by measuring the gap between the correctness obtained on the annotated sentences (i.e. the upper limit the ASR system can reach) and the correctness on the original sentences. Such gap, which decreases along with the grade level, appears to be highly and significantly correlated (see Table 1) with the reference error shown in Fig. 3, being the latter calculated independently on the basis of the edit distance between the annotated sentences and the original sentences.

V. Conclusion

The preliminary ASR system seems to be able to estimate the decoding accuracy of the children and to approximate the reference accuracy calculated on the gold dataset (see Fig. 2 and 3). Nonetheless, the accuracy of the ASR system itself is still poor, especially for young readers (see correctness on 3rd graders in top-left pane of Fig. 1). The improvement of the quality of the sentence annotation together with the creation of a larger gold dataset will help to fill such gap.

Moreover, the next objective consists in estimating, precisely for the words to which the model associates a high level of uncertainty, the sequence of phonemes actually pronounced by the child. This will allow for the automation of the procedure for evaluating the correctness of the decoding of the reading aloud trials. This procedure, for each of the 419 reading trials, was performed manually and these data will constitute a useful benchmark for the automatic analysis system.

A detailed analysis of decoding errors, with particular attention to those words to which the model associates a high level of uncertainty, will be integrated into the ReadLet platform to support professionals to assess the level of reading skills reached by the child, and decide which intervention programmes and measures are most appropriate.

Acknowledgments

This work was supported by the Swiss grant "AEREST: An Ecological Reading Efficiency Screening Tool" (2017-2020) funded by the Department of Teaching and Learning of the University of Applied Sciences and Arts of Southern Switzerland (SUPSI), by the Italian project "(Bio-)computational models of language usage" (2018-) funded by the Italian National Research Council (DUS.AD016.075.004, ILC-CNR), and by the PRIN grant 2017W8HFRX "ReadLet: reading to understand. An ICT-driven, large-scale investigation of early grade children's reading strategies" (2020-2022), from the Italian Ministry of University and Research.

References

[1] Stephen K. Reed. *Cognition. Theories and Applications.* Wadsworth Cengage Learning, Belmont, California (USA), 2012.

[2] OECD. Assessment and Analytical Framework. Technical report, Organisation for Economic Co-operation and Development (OECD), Paris (France), 2019.

[3] Jorge Proença, Carla Lopes, Michael Tjalve, Andreas Stolcke, Sara Candeias, and Fernando Perdigão. Automatic Evaluation of Reading Aloud Performance in Children. *Speech Communication*, 94, 2017.

[4] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5).* American Psychiatric Association, Washington DC (USA), 2013.

[5] Marcello Ferro, Sara Giulivi, and Claudia Cappa. The AEREST Reading Database. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *7th Italian Conference on Computational Linguistics (CLIC-IT'20)*, Torino (Italy), 2020. aAccademia University Press.

[6] Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Ouaphae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. ReadLet: Reading for Understanding. In *IEEE 5th International Congress on Information Science and Technology (CiSt'18)*, pages 1–6, 2018.

[7] Federico Albano Leoni, Francesco Cutugno, Renata Savy, and Valentina Caniparoli. Corpora e lessici di italiano parlato e scritto (CLIPS). `http://www.clips.unina.it/`, 2004. Last accessed on September 9th, 2021.

[8] Dan Su, Xihong Wu, and Lei Xu. Gmm-hmm acoustic model training by a two level procedure with gaussian components determined by automatic model selection. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4890–4893, 2010.

[9] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[10] Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul. A compact model for speaker-adaptive training. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 2, pages 1137–1140. IEEE, 1996.

[11] R. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 2:661–664 vol.2, 1998.