1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), pages 20–26
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Automatically Assess Children's Reading Skills

Ornella Mich, Nadia Mana, Roberto Gretter, Marco Matassoni, Daniele Falavigna
Fondazione Bruno Kessler (FBK), Trento, Italy
{mich, mana, gretter, matasso, falavi}@fbk.eu

Abstract

Assessing reading skills is an important task teachers have to perform at the beginning of a new scholastic year to evaluate the starting level of the class and properly plan next learning activities. Digital tools based on automatic speech recognition (ASR) may be really useful to support teachers in this task, currently very time consuming and prone to human errors. This paper presents a web application for automatically assessing fluency and accuracy of oral reading in children attending Italian primary and lower secondary schools. Our system, based on ASR technology, implements the Cornoldi's MT battery, which is a well-known Italian test to assess reading skills. The front-end of the system has been designed following the participatory design approach by involving end users from the beginning of the creation process. Teachers may use our system to both test student's reading skills and monitor their performance over time. In fact, the system offers an effective graphical visualization of the assessment results for both individual students and entire class. The paper also presents the results of a pilot study to evaluate the system usability with teachers.

Keywords: reading skills, reading assessment, language learning, automatic speech recognition, children's speech recognition

1. Introduction

Assessing reading skills is one of the important tasks that teachers usually perform at the beginning of the scholastic year to have all the information they need to build an overview of the students' reading level and consequently plan effective lessons. This assessment should also be repeated at regular intervals during the scholastic term in order to monitor students' progress and, when necessary, to reformulate the work plan, including specific exercises to strengthen the students' skills and overcome possible difficulties. One of the most well-known standardized tests used in Italy to assess reading skills is based on the MT battery (Cornoldi et al., 1998), which measures the reading fluency, accuracy and comprehension. If the comprehension test can be simultaneously administered to all students of a class, the fluency and accuracy tests must instead be individually administered, in a quiet room: the student is invited to read aloud the piece as best as he/she can, whereas the examiner times and marks the errors on a specific paper sheet. Although it would be desirable to have several evaluation moments during the scholastic term, since this activity is very time consuming, this aspect prevents to regularly repeat the assessment. Furthermore, this activity is also subject to human errors. For these reasons, a digital tool supporting teachers in the MT battery administration seems to be really helpful.

The paper presents a web application for automatically assessing the fluency and the accuracy of oral reading in children attending the primary and lower secondary school. A first prototype of this system was described in a previous paper (Artuso et al., 2017). Here, we will present an advanced version of it, especially focusing on the design of its front-end. We will also describe the new functionalities that support teachers in quickly evaluate reading skills of an entire group of students. Furthermore, the paper presents the results of a pilot study, carried out with teachers, to evaluate the system usability.

This paper is organized as follows: Section 2 reports on research studies related to the paper's topic, whereas Section 3 describes the whole architecture of the system, giving some details about (a) the server side and (b) the client side. Finally, Section 4 draws some conclusions by highlighting benefits and limitations of the system, and presenting directions for future work.

2. Related works

Many studies have demonstrated the effectiveness of technology supporting children's learning by strengthening and enhancing a variety of skills, including those related to reading accuracy, speed, fluency and comprehension (Dynarski et al., 2007; Kamil, 2012). In the last decades, useful applications have been developed to support the reading process through automatic assessment of oral reading by estimating reading errors (Mostow et al., 1993), (dis)fluency (Bolanos et al., 2013), or mispronunciations (Black et al., 2010). Most of these applications are based on automatic speech recognition (ASR) technology. Indeed, the recent advances in the ASR field by means of new hybrid Deep Neural Network–Hidden Markov Models (DNN-HMMs) (Serizel and Giuliani, 2016), trained on large children spoken language corpora or originally developed for adult speech and then adapted to children speech (Serizel and Giuliani, 2014; Giuliani and Babaali, 2015; Liao et al., 2015), have made possible significant improvements of ASR algorithms and fostered the spread of technology based on automatic recognition of children speech for computer-assisted language learning.

The adoption of this technology is fostered by a design process based on a participatory approach (Schuler and Namioka, 1993), where target users (children, parents or teachers) are actively involved in the development stage starting from the beginning. Following this approach, user requirements as well as user needs and expectations are investigated and collected by focus groups, brainstorming meetings, interviews or questionnaires. The gathered information is analyzed and a first draft of the graphical interface is usually elaborated in the form of mock-ups, discussed and commented with the end users in order to collect feed-

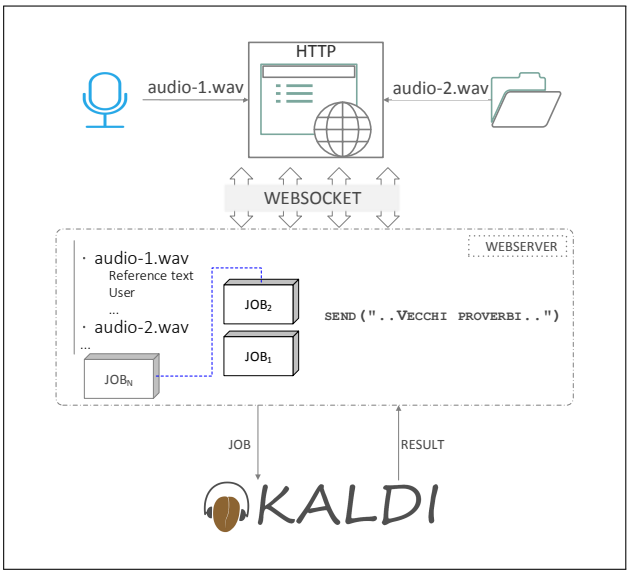

Figure 1: System overview.

back before the implementation stage. Finally, the system is usually tested and evaluated firstly by User Experience (UX) experts, and secondly by target users in order to assess its usability and accessibility (Nielsen, 2012).

3. System architecture

Our system is a web application based on an architecture formed by several modules distributed along both client and server sides. This architecture is illustrated in Figure 1.
On the client side, a web browser acquires audio files either directly from the microphone of the device (PC, laptop, mobile), or uploading them from the file system.
The collected audio files are then sent to the web server and here processed, i.e. compared to the reference reading. The resulting outputs are sent back to the client side, where they are visualized on a web page.
In the following, we briefly describe the server, which is the technological core of our system, and the client i.e. the front-end, which is the graphical interface between the technological core of our system and the user.

3.1. The server

The aim of the server is to process the audio file(s) sent from the client, and return the results to the client, which will visualize them. More specifically, first the server has to perform ASR on the incoming audio, then to compare the text automatically obtained by means of ASR algorithms with the expected text, and find out errors in reading, in particular those concerning speed and accuracy. The results of the comparison are also stored in a database, which allows teachers to create personalized visualizations of the data, for example data aggregated by class, novel, date, individual student, etc.
Our server is built up with the *Node.js*® framework[1]. The audio files acquired by means of the client, as explained in Section 3.2., are first transcribed by means of the ASR module which is based on the KALDI toolkit (Povey et al., 2011), an open source software toolkit largely used to develop state-of-the-art ASR systems for a variety of applications. Since the recorded audio files are related to a predefined set of texts, the automatic transcription related to a submitted job is then aligned with the reference transcription of the corresponding audio recording.

3.1.1. Acoustic models

The training corpus used by the ASR KALDI (Serizel and Giuliani, 2014; Giuliani and Babaali, 2015) consists of clean read speech from Italian children distributed by school grade, from grade 2 through grade 8, i.e. approximately aged from 7 to 13 years.
The training set was built by involving 115 children, each of whom was asked to read 58 or 65 sentences selected from digital texts of children's literature, appropriate for his/her school grade. Each speaker read a different set of sentences, including also 5-8 phonetically reach sentences. The number of utterances in the training set is 7,020 whereas their total duration is 7h:16m.
First, triphone hidden Markov models (HMMs) with gaussian mixture model (GMM) output densities are trained and used to align acoustic observations with tied HMMs states, obtained by means of a phonetic decision tree. Then, a deep neural network (DNN) with output nodes associated to tied HMMs states is trained using the resulting alignment. Acoustic observations are obtained from an eleven frames context window of features (5 frames at each side of the current frame).
Outputs of hidden layers are transformed by sigmoid functions, while softmax normalization is applied to the output layer. The DNN has 4 hidden layers each with 1536 neurons and 2410 output nodes (i.e. the same number of HMMs tied states). See (Artuso et al., 2017) for more details related to both acoustic modeling and decoding process.

3.1.2. Language models

To train the language models used in the ASR system we took advantage from the fact that the texts read by the pupils are those of predefined novels, and therefore known.
To both develop the ASR system and measure its performance, we have considered four different Italian novels, namely *I sette re di Roma* (The seven kings of Rome), *Vecchi proverbi* (Old proverbs), *La botte piena e la botte vuota* (The full barrel and the empty barrel), *I sovrani etruschi* (The Etruscan kings), taken from the Cornoldi's MT battery (Cornoldi et al., 1998), specifically designed and validated by experts to evaluate children's reading skills. A corpus of twenty readings was built by recording children (9 female and 11 male, aged 8-12 yeas) while reading aloud these novels. This corpus was used as testing set to assess the performance of the developed ASR system. Here below, we will give some details of the different language models (LM)s employed, while the reader is addressed to (Artuso et al., 2017) for examining the related achieved results more in details. The texts of all the four novels mentioned above were first normalized by: *a)* removing the punctuation, *b)* expanding numbers and acronyms and *c)* reducing all words to lowercase. Then the following three different 3-gram LMs were trained on the resulting text data, using the IRSTLM open source toolkit (Federico et al., 2008):

[1]https://node.js.org

- Text To Read (TTR). The text training data are the reference texts of the novels, i.e. no attempts to train a reading error model is carried out.
- Automatic Error Model (AEM). The TTR data set is augmented with words formed by syllables obtained from the word beginnings (e.g., *bottiglia* – lit. bottle – generates *bot-* and *botti-*). With this approach we try to simulate false starts.
- Leave One Out (LOO). Both TTR and AEM text data are augmented with "exact" manual transcriptions of the sentences read by the pupil, so that a real reading error model can be trained. In this way the error model can account for non predictable reading errors leading to non-words, like for example mispronunciations of uncommon words or names (for instance *Tarquinio Prisco* often becomes *Tarquinio Parisco*, *proverbio* becomes *provervio*, etc.).

Table 1 shows some samples of the texts used to train the different LMs described above. The total number of words in the four stories is 606, the number of unigrams, bigrams and trigrams resulting after LM training on the TTR data set is: 332, 594 and 12, respectively.

3.2. The front-end

The structure of the front-end side of the second version of our system, i.e. the client, has been re-designed following the participatory design approach (Schuler and Namioka, 1993).
The system's designers involved end users - teachers - from the beginning of their work, organizing focus groups and brainstorming sessions with them to gather their needs and expectations. Pilot studies were performed to test the system between a process step and another.
The client is organized in three main parts: the acquisition page (Figure 3), the visualization of the assessment results of a single student (Figure 4), and the visualization of the assessment results of an entire class (Figures 5 and 6).

Table 1: Texts used to train the LMs. Pronunciation errors are highlighted in bold.

TTR
per la sorpresa e l' amarezza il vecchio proverbio . . .
AEM
pe- per la **so-** sorpresa e l' **ama-** amarezza il **ve-** vecchio **pro-** proverbio . . . per la **sorpre-** sorpresa e l' **amare-** amarezza il vecchio **prove-** proverbio . . .
LOO
per la sorpresa **e l' amarezz-** e l' amarezza **del** vecchio proverbio . . . per la sorpresa e l' amarezza il vecchio **provervio** . . . per la **s-** sorpresa e l' amarezza il vecchio proverbio . . . per la sorpresa e l' amarezza il vecchio proverbio . . . per la sorpresa e l' **armarezza** il vecchio proverbio . . . . . .

Figure 2: Material created and discussed during the focus group with teachers.

Before illustrating in detail each one of the client's parts, as an example of user involvement in the design process, we will describe one of the focus groups we performed with end users after the implementation of the first version of our system (Artuso et al., 2017) in order to collect information to design a better version.

3.2.1. Participatory design

Seven teachers coming from three different elementary schools in our area were involved in the focus group organized to discuss the first version of our system and find out its weaknesses and strengths, to be overcome and emphasized respectively.
The involved teachers were invited to discuss the following topics: (1) the MT battery, (2) the reading aloud practice, (3) possible new functionalities to be added to improve the system, and (4) what are the potentials of our system. For each of the above topics, the teachers first individually worked writing their thoughts on post-its (Figure 2) and then discussing them in group, chaired by two researchers.
Concerning the MT battery, the teachers affirmed that they usually perform the test individually, outside of the classroom, and use it to measure the reading fluency, whereas they globally evaluate the student considering not only the result of the test but also considering the individual progress during the previous scholastic years. The involved teachers also highlighted the fact that the novels proposed by the MT battery are easier than those proposed in the current school textbooks and also than those used in the official national screening - Prove INVALSI[2].
Concerning the reading aloud practice, the involved teachers said it is an important activity: they usually invite students to train this skill at home and then they evaluate the students with reading aloud sessions at school.
Concerning the first version of our system, after working individually with it, the teachers suggested some improvements: (1) adding the possibility of using it in the classroom, where each student has his/her computer and each one can perform the reading test in parallel with other students, because this would allow to save a lot of time; however, this function implies that the system is able to capture

[2]https://www.invalsi.it/

only the audio recorded more closely to the microphone and filter out any noise, which is technically difficult; (2) they suggested to implement a version of the system running on tablet-PCs because they affirm that it is easier for students reading on the display of tablet PCs than on that of desktop PCs; (3) they would like having the possibility of printing the results of the audio elaboration performed by our system to discuss them with other class teachers, as well as with pupils' parents; (4) teachers suggested to also find out the missing pauses and not only highlight those that are too long; (5) teachers would prefer not having a global score including both fluency and accuracy, but having two separated scores.

Concerning the potentials of our application, the teachers stated that it is really interesting because (a) it allows an objective evaluation, (b) it can be used more often than the paper version, (c) it is useful to have a view of the reading skill progress over time. They proposed to also add the possibility of evaluating the reading comprehension skills.

After analysing the results of this focus group, our software programmers implemented the request (5) - having two separated scores and the request (3) - printing the results. The other requests are considered as future work.

3.2.2. The acquisition page

At the beginning of an assessment session, the teacher inserts the name of the student involved in the test and the class he/she is attending; then, the teacher selects the title of the novel on which the student is evaluated. Now, the system is ready to receive the audio file to be elaborated. Concerning this point, our system may work in two different ways: (a) the teacher first records offline the oral reading of the student and then uploads the audio file using the arrow on the right part of the page (see Figure 3 on the top-right); (b) the student reads real time the chosen novel using the PC microphone; in this case, the teacher clicks on the microphone icon (see Figure 3 on the top-left) to start the audio recording. When the system has acquired the entire file (case (a)) or the student has finished to read the text (case (b)), the teacher listens to the recorded audio by clicking on the play button. If he/she is satisfied with it, he/she lets start the audio processing by clicking on the button "TRASCRIVI" ("Transcribe") to launch the ASR algorithms and perform the automatic assessment. Otherwise, the reading can be recorded again.

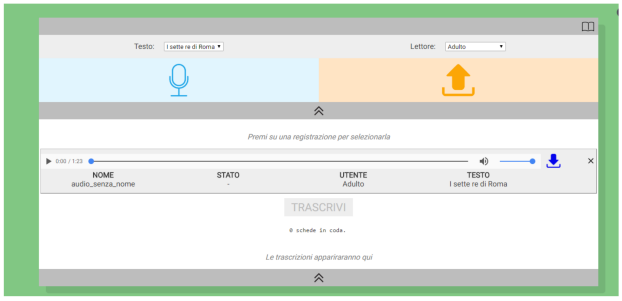

Figure 3: How to insert new registrations.

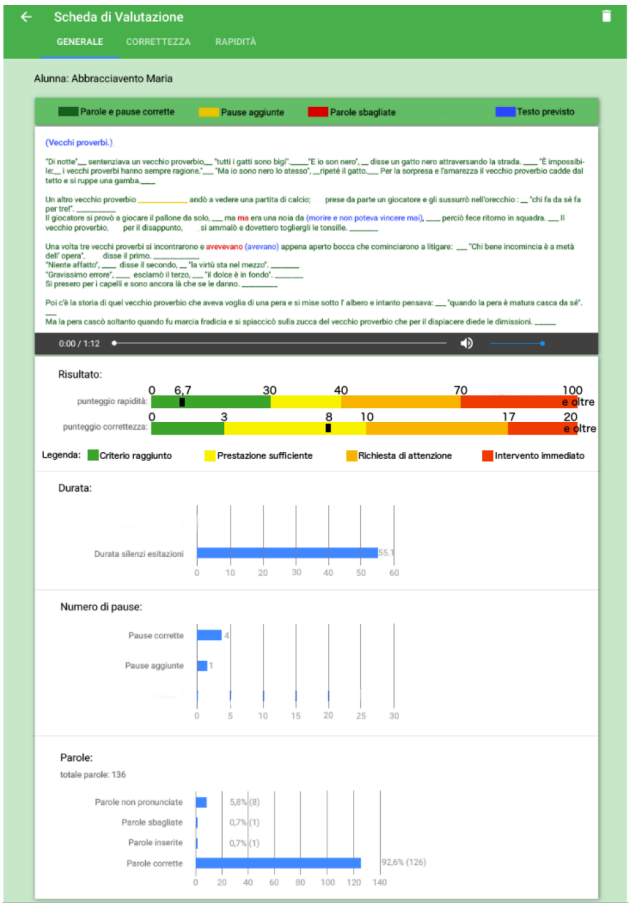

Figure 4: Assessment results of a single student.

3.2.3. Visualization of a student's assessment

At the end of the audio processing made by the ASR on the server-side, our system opens a web page as that shown in Figure 4, which represents the output of our system for a student attending the fifth class of the primary school and reading the novel "I vecchi proverbi". The web page is divided in three main areas: on the top side of the page, the transcription of the read text is visualized; then, there is a box where the scores are visualized, and on the third area, three charts are visualized, reporting results statistics.

The text area Here, words and pauses, represented by means of underscores, are visualized in different colors, chosen by following the rules of the color psychology (Elliot and Maier, 2014), as explained in the following: the green color is used to indicate both the words correctly read and the pauses correctly done; the yellow color is used to indicate the pauses added where not necessary; the red color is used to indicate those words that are not correctly read; the blue color is used to indicate the skipped words, i.e. the words present in the text but not read by the student. This kind of visualization makes the whole area a sort of *picture* of the reading: the teacher has an immediate feeling of the student's performance, without listening the recording.

Individual Scores Our system assesses the reading skills computing the two scores proposed by Cornoldi et al. (1998): (1) speed of reading and (2) accuracy. The speed

Table 2: Schema for placing scores in reading the text *Old Proverbs* (see Figure 4), used to assess students attending the fifth class of the primary school.

	fullness criterion reached	sufficient performance against the criterion	attention required	immediate intervention request
Speed (in cs)	< 31	$31 - 40$	$41 - 70$	> 70
Accuracy (in #errors)	$0 - 3$	$4 - 10$	$11 - 17$	> 18

of reading is computed as the total amount of hundredth of seconds spent by the reader to complete the entire reading divided by the number of read syllables. The accuracy is associated to the total number of errors. An error is the missing of a (group of) syllable(s) or a word, the adding of a (group of) syllable(s) or a word, a pause longer than five seconds, the wrong reading of a syllable.

Cornoldi et al. (1998) measure the reading skills according to four levels: (1) fullness criterion reached, (2) sufficient performance against the criterion, (3) attention required and (4) immediate intervention request. In the case of the text reported in Figure 4, the values associated to each level are reported in Table 2. These values, depending on the number of words and syllables presented in a text, change from text to text.

Individual Statistics In this area, three charts summarize some information about the text's pauses (duration, number of correct ones, number of missed ones) and words (the number of those correctly read, those wrong read, the missed ones). Having a graphical representation of these information helps teachers quickly have highlighted the aspects on which the student is struggling: if the timing, i.e. the pauses, or the spelling.

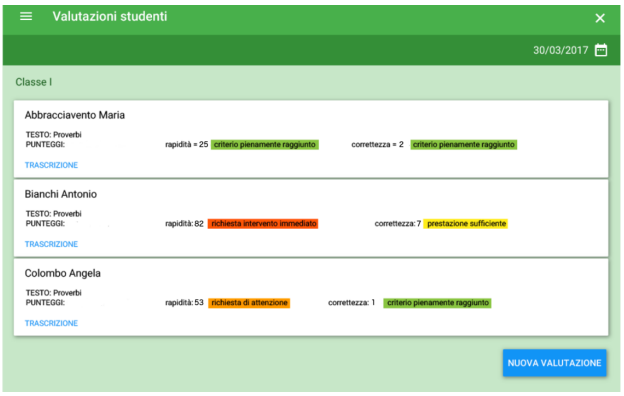

Figure 5: Assessment results of a class.

3.2.4. Visualization of a class's assessment

The two pages summarizing the results of the entire class (Figures 5 and 6) complete the visualization of the outputs of our system. The first summary page reports the single scores of each student in the class, scores related to the reading done in the specific day, which is selected on the calendar that the teacher can open clicking on the icon at the top right of the page (Figure 5). Clicking on the button "Trascrizione", the teacher can directly go to the result page of a specific student (Figure 4).

The second summary page (Figure 6) reports three types of charts: the first one visualizes the mean of the speed and accuracy of the entire class over time (Figure 6 on the top); the second one details the scores of all the students in the class, for a specific day (Figure 6 on the middle and on the bottom); the third one reports the scores of a single student over time.

3.2.5. Usability evaluation with end users

In order to evaluate the usability, meant as ease of access and use (Nielsen, 2012), of our system, we conducted a pilot study by involving three Primary school teachers who teach Italian. One, 28 year old, daily accesses Internet and has a medium level of digital skills, whereas the other two, aged 53 and 56 respectively, are less tech savvy and have a low level of digital skills. After receiving a short description of the application and of the aim of the study, the participants were asked to individually perform the following four tasks: 1. upload a new audio file from the local file system; 2. record a new audio file on the spot by using the PC microphone; 3. launch the automatic transcription and assessment process; 4. search for one of the past transcrip-

Figure 6: Monitoring of a class over time.

tions and check it.
The participants were observed during the tasks in order to: (a) check if and how they could complete the tasks, (b) see which difficulties they met, and (c) collect any comment during the task performing.
The first participant (the youngest and most tech savvy one) quickly completed all the proposed tasks, without any particular difficulty. The other two participants performed without problems tasks 1 and 4, but they both had an hesitation on the task 2 and the third one needed the help of the observer to complete the task 3.
At the end of the experimental session, the participants were interviewed to collect appreciations, criticisms and any suggestion useful to improve the system. In particular, it was explicitly asked (1) do you think that this application would be useful for your job? and (2) what would you improve?
All three participants really appreciated the application and positively replied to the first question. Only one participant added that it would be more useful if the application could work without the Internet connection. Regarding the second question: two participants stated that they would like to have the possibility of loading and processing more than one audio file at a time, and one participant asked for a search bar, supporting a quick search among all the stored transcriptions by student's name.
Given the findings of this pilot study, both the suggestion about the multiple file loading and that about the quick search by name were implemented in the current version of our system, whereas we are investigating the possibility of also making a stand-alone version.

4. Conclusion

In this paper we have presented a web application for automatically assessing the oral reading skills of children. Given audio recordings of oral readings, the system applies ASR algorithms and automatically estimates reading errors, disfluency, hesitations. The system aims to support teachers in assessing reading skills in their students. Starting from a first version presented in (Artuso et al., 2017) the front-end side of the system has been completely redesigned by following a participatory approach that involved a group of teachers both in focus groups and in pilot studies.
At the moment, the system has two main limitations: (a) it only estimates reading accuracy and speed but not comprehension and (b) it does not work on any text but on a set of pre-defined and pre-processed texts.
However, the system offers several advantages by: (a) speeding up the assessment based on the tests of MT battery (Cornoldi et al., 1998), (b) preventing humans errors in timing the task and marking errors, (c) memorizing in the server's database more assessment sessions along the year, so to allow teachers to better monitor and compare students' performance over time, (d) giving details on the errors, helpful for the teacher to suggest specific exercises to overcome students' difficulties, (e) both individual and class monitoring over time, (f) giving a quick visual overview of the errors by means of an effective graphical visualization.
The next steps will be to: 1) add the automatic reading comprehension assessment, and 2) carry out a massive evaluation of the system with teachers of several schools by assessing (a) the accuracy of ASR algorithms and therefore the system precision compared to the humans, (b) the efficiency in terms of time-saving, and (c) the usability of the graphical interface. Future work will also include the design and implementation of (a) a version for students for doing exercises with self-assessment in order to consolidate the reading skills, and (b) an enhanced version of ASR models able to process audio files of any text, instead of a pre-defined set of texts. Finally, in order to make the automatic task even more efficient, we are also going to explore the feasibility of performing parallel assessments in a noise environment such as a classroom where all students are in front of his/her computer, each one performing the reading test in parallel with other students. That means to try to face the limitations due to noisy recordings and to overcome the technical difficulties negatively impacting on the speech recognition accuracy by trying to use appropriate filtering.

5. Acknowledgements

The authors would like to thank all the children who contributed to the corpus of oral readings, as well as all the teachers who participated to the design and experimental sessions. Special thanks also go to Kaleidoscopio Social Cooperative and the school directors for their valuable collaboration in finding participants for our data collections and testing.

6. Bibliographical References

Artuso, S., Cristoforetti, L., Falavigna, D., Gretter, R., Mana, N., and Schiavo, G. (2017). A system for asessing children readings as school. In *SLaTE*, pages 115–120.

Black, M. P., Tepperman, J., and Narayanan, S. S. (2010). Automatic prediction of children's reading ability for high-level literacy assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):1015–1028.

Bolanos, D., Cole, R. A., Ward, W. H., Tindal, G. A., Schwanenflugel, P. J., and Kuhn, M. R. (2013). Automatic assessment of expressive oral reading. *Speech Communication*, 55(2):221–236.

Cornoldi, C., Colpo, G., and Gruppo, M. (1998). Nuove prove di lettura mt. *Giunti OS*.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., Means, B., Murphy, R., Penuel, W., Javitz, H., et al. (2007). Effectiveness of reading and mathematics software products: Findings from the first student cohort. IES Report - hal-00190019.

Elliot, A. J. and Maier, M. A. (2014). Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, 65:95–120.

Federico, M., Bertoldi, N., and Cettolo, M. (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proc. of Interspeech*, pages 1618–1621, Brisbane, Australia, September.

Giuliani, D. and Babaali, B. (2015). Large Vocabulary Children's Speech Recognition with DNN-HMM and

SGMM Acoustic Modeling. In *Proc. of Interspeech*, pages 1635–1639, Dresden (Germany), September.

Kamil, M. L., (2012). *Current and historical perspectives on reading research and instruction*, pages 161–188. American Psychological Association.

Liao, H., Pundak, G., Siohan, O., Carroll, M., Coccaro, N., Jiang, Q.-M., Sainath, T. N., Senior, A., Beaufays, F., and Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. In *Interspeech*.

Mostow, J., Hauptmann, A. G., Chase, L. L., and Roth, S. (1993). Towards a reading coach that listens: Automated detection of oral reading errors. In *AAAI*, pages 392–397.

Nielsen, J. (2012). Usability 101: Introduction to usability (2012). *URL: http://www.nngroup. com/articles/usability-101-introduction-to-usability/[Accessed November 2016]*, 9:35.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Schuler, D. and Namioka, A. (1993). *Participatory design: Principles and practices*. CRC Press.

Serizel, R. and Giuliani, D. (2014). Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In *Proc. of IEEE SLT Workshop*, South Lake Tahoe, (California and Nevada), December, 7-10.

Serizel, R. and Giuliani, D. (2016). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering*, FirstView:1–26, 7.