7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

A SYSTEM FOR ASSESSING CHILDREN READINGS AT SCHOOL

Stefano Artuso[(1,2)], *Luca Cristoforetti*[(2)], *Daniele Falavigna*[(2)], *Roberto Gretter*[(2)], *Nadia Mana*[(2)], *Gianluca Schiavo*[(2)]

[(1)] Pedius Srl, via Boezio 4C, Roma (Italy)
[(2)] Center for Information and Communication Technology
Fondazione Bruno Kessler, via Sommarive 18, Trento (Italy)
{artuso,cristofo,gretter,mana,gschiavo,falavi}@fbk.eu

Abstract

In this paper we describe a system for analyzing the reading errors made by children of the primary and middle schools. To assess the reading skills of children in terms of reading accuracy and speed, a standard reading achievement test, developed by educational psychologists and named "Prove MT" (MT reading test), is used in the Italian schools. This test is based on a set of texts specific for different ages, from 7 to 13 years old. At present, during the test, children are asked to read aloud short stories, while teachers manually write down the reading errors on a sheet and then compute a total score based on several measures, such as duration of the whole reading, number of read syllables per second, number and type of errors, etc. The system we have developed is aimed to support the teachers in this task by automatically detecting the reading errors and estimating the needed measures. To do this we use an automatic speech-to-text transcription system that employs a language model (LM) trained over the texts containing the stories to read. In addition, we embed in the LM an error model that allows to take into account typical reading errors, mostly consisting in pronunciation errors, substitutions of syllables or words, word truncation, etc. To evaluate the performance of our system we collected 20 audio recordings, uttered by 8-13 years old children, reading a novel belonging to "Prove MT" set. It is worth mentioning that the error model proposed in this paper for assessing the reading capabilities of children performs closely to an "oracle" error model obtained from manual transcriptions of the readings themselves.

Index Terms: language learning, children's speech recognition, reading assessment.

1. Introduction

The usage of automatic speech recognition (ASR) systems in language learning applications is not a novelty, although in the past most of the research focused on the development of approaches and tools to aid the learning of the second language [1, 2, 3, 4, 5]. These studies and research resulted into a large number of current commercial products, implementing software tools for computer-assisted language learning (CALL) and assessment, that extensively use ASR technology, mostly for detecting pronunciation errors and for giving feedback to the learner about the quality of her/his pronunciation in the target language[1]. The usage of these products has been largely debated in the language teaching community giving rise, especially in the past, to skepticism [6, 7] by language teachers to accept the results furnished by ASR systems. However, it is a matter of fact that, due to the recent impressive improvement of speech recognition performance on a large set of languages, ASR has become a strategic component in quite all CALL products. Finally, the huge increase in the usage of the Internet and Web based applications has determined an explosion of online courses for second language (L2) learning, even tuned for specific domains (e.g. business, retail, science, etc), that employ ASR as an essential feature.

Despite the large diffusion of the above mentioned L2 CALL tools, little work has been done till now for developing ASR tools for helping language learning by children at school. This is probably due to the fact that commercial ASR products do not perform well with children voice. In fact, it is well known that spectral and temporal characteristics of children's speech are highly influenced by the anatomical, physiological and developmental changes that occur during the growth and are hence different from those of adult speakers [8, 9, 10]. Therefore, when an ASR system, trained on adults' speech, is employed to recognize children's speech, performance decreases drastically, especially for younger children [11, 12, 13, 14, 15, 16, 17, 18, 19].

To compensate for this behavior, assuming having at disposal a small amount of data for training an ASR system for children, we proposed in the past some approaches [18, 20] for adapting the acoustic models of an ASR system trained on adult data to children's speech. In these papers we demonstrated that the acoustic hidden Markov models (HMMs) adapted to children's voices are also effective for large vocabulary speech recognition applications.

Starting from that experience we decided to investigate an application of children's ASR in the language learning domain. More precisely, we developed a tool for both analyzing and assessing the reading errors made by children of Italian primary and middle schools when they are examined for "Prove MT"(MT Reading Test)[21]. This is a standardized reading achievement test created by experts on diagnosis and treatment of learning disorders through a study involving over 8000 students from schools of all over Italy. The test is commonly used for the assessment of reading speed and accuracy in different school grades. Similar text-reading tests in other languages are the GORT-5 [22] for English, the Alouette test [23] for French and the SLRT II [24] for German. MT Reading Test procedure involves asking children, aged between 7 to 13, to read aloud short stories, each formed by some hundreds of words. During the readings, the teacher manually takes note of both reading errors and reading time. Basically, two parameters are computed from teacher's notes: (i) reading speed, measured in syllables per second, and (ii) accuracy of text reading, measured as the number of errors (e.g.: skipped lines, replaces, omits, adds or reverses letters, makes up words, pauses and hesitations). These

[1] Visit https://en.wikipedia.org/wiki/List_of_language_self-study_programs to view a list of CALL products.

10.21437/SLaTE.2017-20

measures are compared to standard reference data and finally used to produce historical records of the examined students, from which to derive statistics specific either of each student or of entire classrooms.

The system we have developed aims to support the work of the teachers both by automatically detecting the reading errors and by estimating the parameters they usually measure during the reading itself. To do this we use an automatic speech-to-text transcription system that employs a LM trained over the texts containing the stories to read. In addition, we embed in the LM an error model that allows to take into account typical reading errors, mostly consisting in words truncation. Finally, we propose a measure, to score each reading, that takes into account both the fluency of the reading itself, basically based on the total duration of the "erroneous" pauses, and the total number of automatically detected errors.

This paper is organized as follows: Section 2 introduces the topic, discussing the main approaches, Section 3 describes the whole architecture of the system focusing on the user interface (see Section 3.1), Section 4 gives the details of the ASR system, experiments and results are reported and discussed in Section 5. Finally, Section 6 draws some conclusions and discusses directions for future work.

2. Related works

Numerous studies have demonstrated the effectiveness of support technology on a variety of skills related to reading, including vocabulary, phonemic awareness, reading fluency, speed and comprehension [25]. Several applications have been developed not only to support the reading process but also to provide automatic assessment of oral reading by estimating errors [26], mispronunciations [27] or reading (dis)fluency [28].

Recently, the use of the hybrid deep neural network (DNN)-HMMs has been shown effective for children's speech recognition [29, 30, 17, 18]. However, a shortage of training data is typical issue when developing ASR systems for children. To cope with this problem, in [18] an age adaptation approach is investigated in which a DNN-HMMs system is first trained on a substantial amount of adults' speech integrated with a small amount (few hours) of children's data. The resulting acoustic model is then adapted to children's voices by exploiting the available children's training data. In a noticeable work [17], a huge amount of children's speech data, more than 1,000 hours, was collected from the Internet and used to train a DNN-HMMs based recognition system with excellent recognition results. Furthermore, it was shown that training an hybrid DNN-HMMs system using balanced large amounts of speech from adults (more than 1,000 hours) and children (more than 1,000 hours) leads to recognition performance equal or better than training specific acoustic models for adults and children. Contrastive experiments were conducted using different neural network models and configurations, confirming these results. However, such large amount of children's speech is rarely available for training purposes especially for languages different than English.

In [20] we described an approach, based on objective function regularization [31], for adapting to children a DNN trained on around 300 hours of voices of adult speakers. We showed its effectiveness even when the size of adaptation data is quite small (i.e. less than one hour speech). Moreover, we observed that the ASR performance does not deteriorate too much also when manual supervision of the adaptation data is not available (i.e. the supervision of the data is generated by an ASR).

In [30] a maximum likelihood linear regression (MLLR)-based speaker adaptive training (SAT) DNN-HMMs system trained on a small amount, 7 hours, of children's speech was shown effective in a large vocabulary continuous speech recognition (LVCSR) task. Similar results were also achieved in experiments reported in both [18] and [20] therefore, for this work, we have decided to use the ASR system described in [20] (see section 4), which was trained on the above mentioned small children data set (around 7 hours of speech). Moreover, since this work is aimed at detecting errors in children readings, given a "small" set of "short" stories (whose texts are known), the size of the dictionary to recognize results quiet small (a few hundreds of words), thus reducing the impact of the quality of the acoustic model on the ASR system performance. Instead, it is crucial for this language learning task to effectively model the errors made by children during their readings, measuring the ability of the whole system to correctly detect them.

3. System architecture

Figure 1 illustrates the whole client-server architecture of the system developed for analyzing children's readings. It is formed by several modules distributed along both client and server sides.

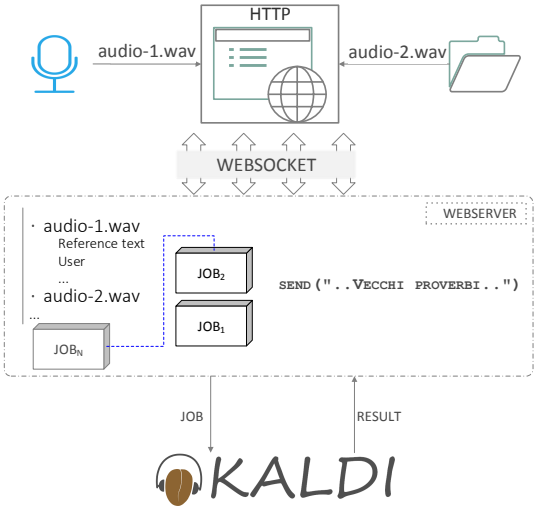

Figure 1: *System overview: audio files are sent to a Node.js® web server which starts the ASR (based on the Kaldi toolkit) system. Jobs containing all audio files and their properties are created and enqueued. When Kaldi completes a transcription task, the webserver handles the result and sends it back to the client side. A full duplex communication between client and server is provided by a websocket.*

On client side, a web browser acquires audio files either from the microphone on the device or reading it from file. Audio files (note that more than one single audio file can be sent to the server) are sent to a web server, built up with *Node.js®* framework [32], through a websocket connection provided by a "Node" module called SOCKET.IO [33]. Over the websocket protocol, client and server exchange messages and data in a standardized and secure way, facilitating real-time data transfer. The *Node.js®* server receives the audio files sent by the user and creates a *job* containing all the audio files and their related information, like the corresponding reference texts. The auto-

matic transcription related to a submitted job, obtained through the KALDI toolkit (see section 4), is aligned with the reference transcription of the corresponding audio recording and an HTML page is produced, as will be explained in section 3.1. This page shows, in a single shot, the detected reading errors, the duration of the erroneous pauses (i.e. the pauses that are not correlated with punctuation) and the score gained in the reading.

3.1. User interface

For a given reading, the system will produce an HTML page containing all the information needed for its evaluation. This information is obtained by means of the following processing steps:

- ASR processing, furnishing a file which contains, in each row, a recognized word with the corresponding time boundaries information;
- alignment between the ASR output and the normalized reference text of the read novel, including timing information;
- computation of two scores: taking into account errors in the word pronunciation and "fluency" in reading (measured using the durations of the pauses), respectively.

The generated HTML page (see Figure 2) is a picture that represents, at a glance, the errors made by the child who read the assigned "Prove MT" text. In the Figure, colors other than blue highlight inserted or missed words while pauses, indicated by underscores, are blue or red depending on whether they correspond to some punctuation or not. In this way, the teacher can have an immediate overview of the reading quality and can easily locate where reading difficulties may have occurred. Furthermore, teachers can listen to the audio and, by clicking over a selected word, they can place the audio pointer to that word. At top of HTML page the scores achieved in the reading are displayed. At present, each reading is assigned an overall score which is the arithmetic mean between two scores, called: "Rhythm Score" (RS) and "Word Score" (WS), respectively. Rhythm Score takes into account: words durations (WD), duration of "good pauses" (GPD, i.e. segments of silences corresponding to punctuation in the text to read) and duration of hesitations (HD, i.e. segments of silences occurring although there is no punctuation to follow in the text). RS is defined as follows:

$$RS = 10 * \frac{WD + GPD}{WD + GPD + HD} \quad (1)$$

Word score takes into account the number of words correctly recognized (CRW) and the number of words wrongly recognized (WRW, which sums both insertions and substitutions errors). WS is defined by the following equation:

$$WS = 10 * \frac{CRW}{CRW + WRW} \quad (2)$$

Absence of hesitations gives $RS = 10$, no recognition errors gives $WS = 10$.

Finally, several statistics can be computed from the whole set of collected records, relatively to the single reader, or in aggregated forms, e.g. referring to: age, gender, classroom, etc.

4. ASR system

The ASR system is based on the KALDI open source software toolkit [34]. The latter, largely used to develop state-of-the-art ASR systems for a variety of applications, integrates hybrid DNN-HMMs in a static decoding network built by means

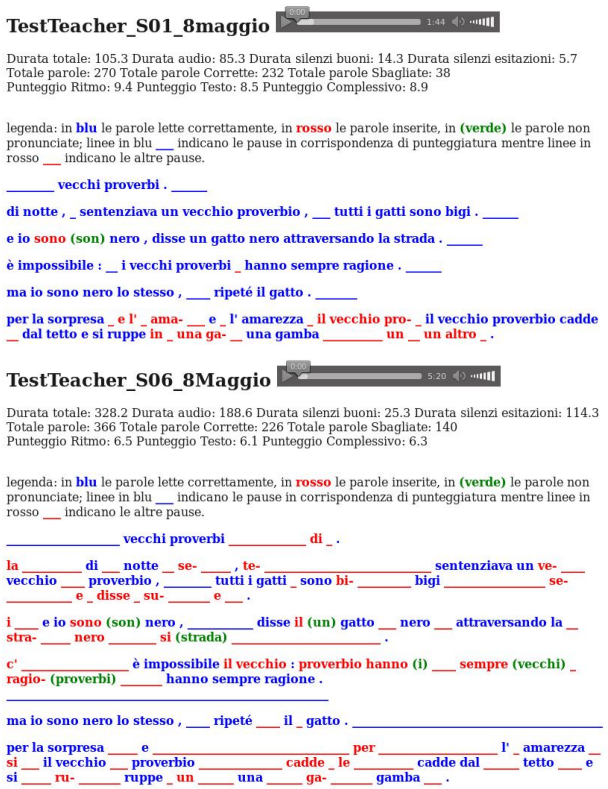

Figure 2: *Output produced by the developed system: the upper Figure corresponds to a "quite good" reading, except the last sentence where there are some hesitations. On the contrary, the lower Figure corresponds to a worse reading: in fact, the number of red pauses and words gives an immediate view of the errors made by the child. Clicking over a blue word locates the audio player to that position, allowing to listen the corresponding part of the recording.*

of finite state transducers. Context dependent HMMs with tied states, speaker adaptive training via MLLR transformations [35], linear discriminant transformation of acoustic observations are some, among the many other features, furnished by the toolkit.

4.1. Acoustic models

For HMMs training, 13 mel-frequency cepstral coefficients (MFCCs) are computed every 10*ms* by using a Hamming window of 25*ms* length. These features are mean/variance normalized on a speaker-by-speaker basis, spliced by +/- 3 frames next to the central frame and projected down to 40 dimensions using linear discriminant analysis (LDA) and a single maximum likelihood linear transformation (MLLT). Then, a single transform is estimated for each training speaker and applied for normalizing features (since the transform is applied to features vectors it is usually named fMLLR) for speaker adaptive training of triphone HMMs.

The training corpus, hereinafter called "ChildIt" [29, 30], consists of clean read speech from Italian children aged from 7 to 13 years, with a mean age of 10 years.

The training set was collected from 115 children, each child read 58 or 65 sentences selected from electronic texts concern-

ing literature for children, depending on his/her grade. Each speaker read a different set of sentences which included, however, a set of phonetically rich sentences (5-8 sentences) which were repeated by several speakers. Speech was acquired at 16 kHz, with 16 bit accuracy, using a Shure SM10A head-worn microphone. The number of utterances in the training set is 7,020, their total duration is 7h:16m.

First, triphone 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 DNN with output nodes univocally associated to tied HMMs states is trained using the resulting alignment. To do this, an eleven frames context window of LDA+MLLT+fMLLR features (5 frames at each side of the current frame) form a 440 dimensional input feature vector for the DNN. This is trained in several stages [36], including: *i*) pretraining consisting in layer-wise training of Restricted Boltzmann Machines (RBM) by Contrastive Divergence algorithm; *ii*) frame classification training based on mini-batch Stochastic Gradient Descent (SGD), optimizing frame cross-entropy; *iii*) sequence discriminative training [37] consisting in SGD with per-sentence updates, optimizing state Minimum Bayes Risk (sMBR). 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).

During decoding the computation of MLLR features is done in two steps. First, a word lattice is produced for each input utterance by using the baseline speaker-independent GMM-HMM. A single MLLR transform for each speaker is then estimated from sufficient statistics collected from word lattices with respect to speaker-adaptively trained triphone HMMs. These transforms are used with SAT triphone HMMs to produce new word lattices. A second set of fMLLR transforms is estimated from new word lattices and combined with the first set of transforms. The resulting transforms are used for features normalization. Finally, output probabilities of HMMs states are computed by dividing the DNN output node posterior probabilities by their corresponding prior probabilities, as explained in [36].

4.2. Language models

As mentioned above, the system we developed for assessing the correctness of readings requires the prior knowledge of the reference texts to read, while the readings errors have to be detected.

In the experiments reported below we are given 4 different Italian novels (namely: "The seven kings of Rome", "Old proverbs", "The full barrel and the empty barrel", "The Etruscan sovereigns"). These novels are read by children during their "Prove MT" trials.

First, the texts of all the 4 stories are normalized, i.e.: words are lowercased, punctuation is removed, numbers and acronyms are expanded - e.g.: VI $\rightarrow$ sesto (sixth); A.C. $\rightarrow$ avanti cristo (Before Christ). Compound words are splitted (e.g. romano-sabini $\rightarrow$ romano sabini), special characters are normalized using the "Latin-1" encoding. The total number of words in the 4 stories is 606, the dictionary size is 332, the number of unigrams, bigrams and trigrams resulting after LM training is: 332, 594 and 12, respectively.

Starting from the normalized texts of the 4 stories, we trained the following three different 3-gram LMs with the IRSTLM open source toolkit [38].

- **Text To Read (TTR)**: only the normalized texts are used to train the LM, without including any error model.
- **Automatic Error Model (AEM)**: the previous document text (TTR) is augmented by corrupting it with an error model which includes all possible word beginnings, on a syllable-like basis, due to false starts. For instance, the words *vecchi* and *proverbi* can give raise to the following false start words: *ve-*, *pro-* and *prove-* that are inserted in the text before the corresponding full words, as shown in Table 1. This approach is a simple, but effective, way to automatically model typical readings errors made by children when they encounter a "not-so-easy" word to read.
- **Leave-One-Out (LOO)**: the previous document text (AEM) is augmented by including in it the manual transcriptions of all the readings made by the children in the evaluation set (see Section 5), except that of the child being evaluated. This is a practical way to insert in the model common but not predictable errors, like for example mispronunciations of uncommon words or names (for instance *Tarquinio Prisco* often becomes *Tarquinio Parisco*, *proverbio* becomes *provervio*, etc).

Table 1 reports some samples of the texts used to train the different LMs.

Table 1: *Texts used to train the LMs. TTR contain the plain reference texts only; AEM expands TTR with automatically generated syllables derived from word beginnings, LOO expands AEM with manual transcriptions of the utterances in the test set, in a leave-one-out fashion. 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 ... ...

5. Experiments and results

Experiments were conducted on a data set of readings containing the recordings of the novel "old proverbs" uttered by 20 children, aged between 8 and 13 years. Note that, although the total number of children involved in our experimentation is 20, not all of them read the 4 previously mentioned novels. All the readings were carefully transcribed manually, annotating: words, words fragments and hesitations. Since the text of the novel "old proverbs" consists of 242 words, the number of running words in the test set results to be $20 \times 242 = 4840$.

Performance was computed using two different references:

- **Ideal Reference Text (IRT)**: it is the normalized text of the novel. This is the reference to consider when preparing the output of the system for the teachers. Comparing the system output against IRT allows to compute a score that can be used to measure the correctness of the reading. This score should be as similar as possible to the one given by the teacher.
- **Manual Transcription Reference (MTR)**: it is the orthographic transcription of the reading, and includes all errors made by the pupil. This is the reference to use to estimate the performance of the ASR system. Typically, a better error model should return a lower word error rate (WER). Note that the MTR texts are the same used to to train the LLO LMs defined in Section 4.2. The total number of running words in MTR texts is 5135.

Table 2 reports the performance achieved on the test set described above, in terms of WER, and using as reference the two previously mentioned transcriptions: IRT and MTR.

The upper two blocks report performance achieved using IRT as reference. The first raw (MTR) gives the WER (together with the percentage of insertions, deletions and substitutions) obtained when the same MTR texts (i.e. the reference manual transcriptions) are aligned with the IRT reference. This value provides a sort of WER oracle (13.90%), since it represents the exact number of reading errors made by children (i.e. we are assuming that the ASR system doesn't make errors). In the second block the 3 LMs defined in section 4.2 are used by the ASR system and allow to score each reading in a completely automatic way. It can be noticed that the proposed error models (both AEM and LOO) give better performance than the baseline LM (TTR, trained without using any error model), and in particular LOO performs closely to the oracle (13.95% versus 13.90%). It is worth noting (see also Table 3) that, in case of hesitation errors, in most cases every LM detects some errors. But while TTR can only insert existing words, AEM and LOO can insert word fragments, often corresponding to the exact error.

Finally, the lower block in Table 2 reports the results when the ASR outputs are aligned with the MTR reference. A perfect ASR system - that knows in advance all possible reading errors and that never makes mistakes - should return 0% WER. In practice, the more accurate the system, the lower the WER obtained. It is important to notice that detecting a word fragment which is similar, but not the same, to the one really pronounced, results in a recognition error. For instance, see hesitation **sog-** in rows 7 of Table 3: it is an error for TTR (recognized as another word **SUD**, *South*) and AEM (recognized as the word fragment **SAGGIA-**); only LOO could detect it correctly as **sog-** because some other children did the same hesitation.

Table 2: *WERs achieved on the test set of children readings using two different reference texts (IRT and MTR). The LMs employed by the ASR system have been trained including (AEM, LOO) or not (TTR) fragment words in the corresponding training sets.*

alignment against IRT - score oracle					
	Words	Subst	Ins	Del	WER:
MTR	4840	5.29%	7.36%	1.26%	13.90%
alignment agains IRT - automatic scores					
TTR	4840	4.05%	8.55%	3.22%	15.83%
AEM	4840	3.00%	9.07%	2.23%	14.30%
LOO	4840	4.21%	8.51%	1.22%	13.95%
alignment agains MTR - goodness of the error model					
TTR	5135	7.93%	2.28%	3.00%	13.20%
AEM	5135	6.97%	2.57%	1.87%	11.41%
LOO	5135	4.73%	2.03%	0.90%	7.65%

Table 3: *Some text samples for a given reading: differences with respect to the IRT are shown in bold, while errors with respect to the MTR are shown in uppercase. TTR cannot insert word fragments, AEM can insert only automatically predicted errors, LOO can also insert already observed errors. Very often, in presence of a pronunciation error, every LM is capable to detect that "something wrong" happened.*

IRT	per la sorpresa e per l' amarezza il vecchio proverbio cadde dal tetto e si ruppe una gamba
MTR	per la sorpresa e **l' ama- e** l' amarezza **il vecchio pro-** il vecchio proverbio cadde dal tetto e si ruppe **in una ga-** una gamba
TTR	*** la sorpresa e **LA MA e** l' amarezza **il vecchio PER** il vecchio proverbio cadde dal tetto e si ruppe **in una** *** una gamba
AEM	per la sorpresa e l' **ama- * *** amarezza **il vecchio pro-** il vecchio proverbio cadde dal tetto e si ruppe ** **una ga-** una gamba
LOO	per la sorpresa e **l' ama- e** l' amarezza **il vecchio pro-** il vecchio proverbio cadde dal tetto e si ruppe **in una ga-** una gamba
IRT	i re etruschi governavano saggiamente portando la civiltà del loro popolo
MTR	i re etruschi governavano **sog-** saggiamente **pa- par-** portando la civiltà del loro popolo
TTR	i re etruschi governavano **SUD** saggiamente **PERA PARTE** portando la civiltà del loro popolo
AEM	i re etruschi governavano **SAGGIA-** saggiamente **pa- PARTE** portando la civiltà del loro popolo
LOO	i re etruschi governavano **sog-** saggiamente **pa- par-** portando la civiltà del loro popolo

6. Conclusions

In this paper we have presented an approach for automatically assessing the reading capabilities of children. The developed system makes use of the time information provided by an ASR system for estimating the durations of the pauses uttered during readings, detecting those corresponding to hesitations. In addition, different LMs have been trained, including in their training data fragment words representing typical reading errors made by children. ASR performance, depending on each LM, has been computed on a manually transcribed test set and effectiveness of the proposed error models has been demonstrated.

Future works include the possibility by the teachers to correct errors made by the automatic system. This will allow both to regenerate the HTML page and present to the pupils/parents a more accurate score, and in the long run to obtain more reliable data to be used to retrain and improve the system itself. Furthermore, a focus group with teachers will be organized to better investigate their expectations from the system in term of possible functionalities to be included, as well as to collect their opinions about the current user interface. Finally, a user study will be conducted to evaluate the interface usability and to assess system performance. These activities will open the way for a real usage of the system in the primary school.

7. References

[1] H. F. et al., "The sri eduspeak system:recognition and pronunciation scoring for language learning," in *Proc. of InSTIL*, Scotland, 2000, pp. 123–128.

[2] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, pp. 95–108, 2000.

[3] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer Assisted Language Learning*, vol. 15, pp. 441–447, 2002.

[4] A. Neri, C. Cucchiarini, and H. Strik, "Automatic speech recognition for second language learning: How and why it actually works," in *Proc. of ICPhS*, Scotland, 2003, pp. 1157–1160.

[5] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech and Language*, vol. 21, pp. 219–230, 2007.

[6] D. Coniam, "Voice recognition software accuracy with second language speakers of english," *SYSYEM*, no. 27, pp. 49–64, 1999.

[7] T. M. Derwing, M. J. Munro, and M. Carbonaro, "Does popular speech recognition software work with esl speech?" *TESOL quarterly*, no. 34, pp. 592–603, 2000.

[8] R. D. Kent and L. L. Forner, "Speech segment durations in sentence recitations by children and adult s," *JPhon*, vol. 8, pp. 157–168, 1980.

[9] S. Lee, A. Potamianos, and S. Narayanan, "Acoustic of children's speech: Developmental changes of temp oral and spectral parameters," vol. 105, no. 3, pp. 1455–1468, March 1999.

[10] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, "Formants of Children, Women and Men: the Effect of Vocal Intensity Variation," vol. 106, no. 3, pp. 1532–1542, September 1999.

[11] J. Wilpon and C. Jacobsen, "A Study of Speech Recognition for Children and Elderly," Atlanta, GA, May 1996, pp. I–349–352.

[12] S. Das, D. Nix, and M. Picheny, "Improvements in Children's Speech Recognition Performance," Seattle,WA, May 1998, pp. 433–436.

[13] Q. Li and M. Russell, "Why is Automatic Recognition of Children's Speech Difficult?" ," in *EUROSPEECH*, Aalborg, Denmark, Sept. 2001.

[14] D. Giuliani and M. Gerosa, "Investigating Recognition of Children Speech," in *Proc. of ICASSP*, vol. 2, Hong Kong, Apr. 2003, pp. 137–40.

[15] A. Potamianos and S. Narayanan, "Robust Recognition of Children's Speech," vol. 11, no. 6, pp. 603–615, Nov. 2003.

[16] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of childrens speech," *Speech Communication*, vol. 49, no. 1011, pp. 847 – 860, 2007.

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

[18] R. Serizel and D. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, vol. FirstView, pp. 1–26, 7 2016.

[19] M. Gerosa, D. Giuliani, and F. Brugnara, "Towards age-independent acoustic modeling," *Speech Communication*, vol. 51, no. 6, pp. 499 – 509, 2009.

[20] D. Falavigna, M. Matassoni, and D. Giuliani, "Dnn adaptation for recognition of children speech through automatic utterance selection," in *Proc. of Spoken Language Technology workshop*, San Diego (CA), US, December 2016.

[21] C. Cornoldi and G. Colpo, *Prove di lettura MT per la scuola elementare-2: 1o e 2o elementare*. Giunti, 2004.

[22] J. L. Wiederholt and B. R. Bryant, *Gray oral reading tests*, 2012.

[23] P. Lefavrais, *Test de lAlouette*. Editions du centre de psychologie appliquée, 1967.

[24] K. Moll and K. Landerl, *SLRT-II: Lese-und Rechtschreibtest; Weiterentwicklung des Salzburger Lese-und Rechtschreibtests (SLRT)*. Huber, 2010.

[25] M. L. Kamil, "Current and historical perspectives on reading research and instruction," *APA educational psychology handbook: Application to teaching and learning*, pp. 161–188, 2012.

[26] J. Mostow, A. G. Hauptmann, L. L. Chase, and S. Roth, "Towards a reading coach that listens: Automated detection of oral reading errors," in *AAAI*, 1993, pp. 392–397.

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

[28] D. Bolanos, R. A. Cole, W. H. Ward, G. A. Tindal, P. J. Schwanenflugel, and M. R. Kuhn, "Automatic assessment of expressive oral reading," *Speech Communication*, vol. 55, no. 2, pp. 221–236, 2013.

[29] R. Serizel and D. Giuliani, "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 2014.

[30] D. Giuliani and B. Babaali, "Large Vocabulary Childrens Speech Recognition with DNN-HMM and SGMM Acoustic Modeling," in *Proc. of Interspeech*, Dresden (Germany), September 2015, pp. 1635–1639.

[31] D. Falavigna, M. Matassoni, S. Jalalvand, M. Negri, and M. Turchi, "DNN adaptation by automatic quality estimation of asr hypotheses," *Computer Speech & Language*, 2016.

[32] "Node.js®," https://nodejs.org/en/.

[33] "Socket.IO," https://socket.io/.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Veselỳ, "The Kaldi Speech Recognition Toolkit," in *Proc. of IEEE ASRU Workshop*, Hawaii (US), December 2011.

[35] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[36] G. Hinton, L. Deng, D. Yu, and Y. Wang, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 9, no. 3, pp. 82–97, 2012.

[37] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative Training of Deep Neural Networks," in *Proc. of Interspeech*, 2011, pp. 2345–2349.

[38] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proc. of Interspeech*, Brisbane, Australia, September 2008, pp. 1618–1621.