Speech Communication 94 (2017) 1–14

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic evaluation of reading aloud performance in children

Jorge Proença[a,b,*], Carla Lopes[a,c], Michael Tjalve[d], Andreas Stolcke[e], Sara Candeias[f], Fernando Perdigão[a,b]

[a] Instituto de Telecomunicações, Coimbra, Portugal
[b] Department of Electrical and Computer Engineering, University of Coimbra, Portugal
[c] Polytechninc Institute of Leiria, Leiria, Portugal
[d] Microsoft & University of Washington, Seattle, WA, USA
[e] Microsoft Research, Mountain View, CA, USA
[f] Microsoft, Lisbon, Portugal

ARTICLE INFO

Article history:
Received 19 December 2016
Revised 27 June 2017
Available online 18 August 2017

Keywords:
Reading level assessment
Child speech
Pseudoword reading
Disfluency detection
Gaussian process regression

ABSTRACT

Evaluating children's reading aloud proficiency is typically a task done by teachers on an individual basis, where reading time and wrong words are marked manually. A computational tool that assists with recording reading tasks, automatically analyzing them and outputting performance related metrics could be a significant help to teachers. Working towards that goal, this work presents an approach to automatically predict the overall reading aloud ability of primary school children by employing automatic speech processing methods. Reading tasks were designed focused on sentences and pseudowords, so as to obtain complementary information from the two distinct assignments. A dataset was collected with recordings of 284 children aged 6–10 years reading in native European Portuguese. The most common disfluencies identified include intra-word pauses, phonetic extensions, false starts, repetitions, and mispronunciations. To automatically detect reading disfluencies, we first target extra events by employing task-specific lattices for decoding that allow syllable-based false starts as well as repetitions of words and sequences of words. Then, mispronunciations are detected based on the log likelihood ratio between the recognized and target words. The opinions of primary school teachers were gathered as ground truth of overall reading aloud performance, who provided 0–5 scores closely related to the expected performance at the end of each grade. To predict these scores, various features were extracted by automatic annotation and regression models were trained. Gaussian process regression proved to be the most successful approach. Feature selection from both sentence and pseudoword tasks give the closest predictions, with a correlation of 0.944 compared to the teachers' grading. Compared to the use of manual annotation, where the best models obtained give a correlation of 0.949, there was a relative decrease of only 0.5% for using automatic annotations to extract features. The error rate of predicted scores relative to ground truth also proved to be smaller than the deviation of evaluators' opinion per child.

1. Introduction

To evaluate the reading aloud ability of primary school children, teachers or tutors usually need to make the effort of providing a level-appropriate reading task to the child, manually take notes for time and accuracy, and calculate a metric such as *correct words per minute*. This 1-on-1 procedure can be very time-consuming, especially if additional performance metrics are desired. Also, manual evaluations are not consistently equal and depend on evaluator bias and experience. An automatic system that can perform these steps accurately would be a great complement to the usual methods and an indispensable tool for teachers that may have classes with up to 30 children. It could also lead to more frequent assessments of a child throughout the school year, and a higher-quality accompaniment of their education. Providing an overall performance score, as opposed to specific metrics and subjective parameters, can give a clear overview of a child's status and can also be beneficial for an analysis of a child's progress over time.

Although this work targets the widespread evaluation of reading of all school children aged 6–10 years, the automatic assessment of reading aloud may also be helpful to detect reading disorders and find specific problems. Furthermore, the same

* Corresponding author at: Department of Electrical and Computer Engineering, University of Coimbra, Portugal.
E-mail addresses: jproenca@co.it.pt (J. Proença), calopes@co.it.pt (C. Lopes), michael.tjalve@microsoft.com (M. Tjalve), andreas.stolcke@microsoft.com (A. Stolcke), v-sacand@microsoft.com (S. Candeias), fp@co.it.pt (F. Perdigão).

http://dx.doi.org/10.1016/j.specom.2017.08.006
0167-6393/

technology and methods are inherently connected to other applications such as automatic reading tutors where, for example, a child's reading is tracked in real-time against the written text and incorrect pronunciations are detected. Some projects aimed to create an automatic reading tutor that follows and analyzes a child's reading, such as LISTEN (Mostow et al., 1994), Tball (Black et al., 2007), SPACE (Duchateau et al., 2009) and FLORA (Bolaños et al., 2011). Other similar applications fall in the area of computer assisted language learning (CALL), where most efforts are for foreign language learning (Abdou et al., 2006; Cincarek et al., 2009) and are targeted to adults or young adults, for whom speech recognition and speech technologies are relatively mature.

It should be emphasized that the current work is concerned with oral reading fluency evaluation, and no effort is made to measure comprehension of what is being read. Nevertheless, there is evidence that oral reading fluency is an indicator of overall reading competence (Fuchs et al., 2001). Oral reading fluency in children is defined as the ability to read text quickly, accurately and with proper expression (Buescu et al., 2015; National Reading Panel, 2000).

To be able to automatically assess the reading aloud performance of children, deviations to the intended correct reading in the form of disfluencies or hesitations must be detected. These are linguistic events which affect the smooth flow of speech, such as repetitions, mispronunciations, cut-off words and corrected false starts (Candeias et al., 2013). There are several known methods to detect disfluencies, such as based on hidden Markov models (HMMs), maximum entropy models, conditional random fields (Liu et al., 2005) and classification and regression trees (Medeiros et al., 2013), though most of these efforts focus on spontaneous speech. Applicability to read speech is not a given since different speaking styles vary in the production of disfluencies (Moniz et al., 2014). Disfluencies in reading have different nuances, and some prior work has targeted the automatic detection of these events in children's reading, with the most relevant contributions described below. Some of the studies mentioned in the following paragraphs also aim to automatically provide an overall reading ability score, closely predicting human evaluation.

Black et al. (2007) aimed to automatically detect disfluencies in isolated word reading tasks. They found that human evaluators rated fluency as importantly as accuracy when judging reading ability. The target of detection was mostly sounding-outs, where a child first reads phoneme by phoneme (which can be whispered) and then reads the complete word. They build HMMs and a grammar structure specialized for disfluencies, capable of detecting partial words and allowing silence or noise between phones. The correct word is compulsorily considered to be pronounced in the final state of the grammar. They achieve 14.9% miss rate and 8.9% false alarm rate for the detection of hesitations, sounding-outs, and whispering. By comparison, in our data, no phoneme by phoneme sounding-out was found. Instead, there are syllable by syllable sounding-outs with possible silence between syllables, which we will address. An extension (Black et al., 2011) aimed to automatically evaluate reading ability and provide a high-level literacy score. Eleven human evaluators of different backgrounds (linguistics, engineering, speech research) rated children's performance in individual word reading tasks with scores from 1 to 7. Using automatically extracted features and a selection of features based on pronunciation, fluency and speech rate, a Pearson correlation of 0.946 was achieved to predict mean evaluator's scores.

Duchateau et al. (2007) also target the reading of isolated words. Based on HMMs, they use a two-layer decoding module, first with phoneme decoding using phoneme-level finite state transducers to allow false starts with partial pronunciations, and then a second lattice allows for repetitions or deletions of words. For the detection of reading errors on word reading, a miss rate of 44% and a false alarm rate of 13% were achieved. For a task of pseudoword reading, they achieve a 26% rate of both misses and false alarms. They evaluate a child's reading ability by the number of correctly read words divided by time spent (same as correct words per minute) and show agreement to human evaluation with Cohen's Kappa (Cohen, 1960) above 0.6 when considering 5 performance classes. In Yilmaz et al. (2014), an extension to the work done in Duchateau et al. (2007) is developed. The new evaluation is on a mixture of word and sentence reading tasks, and the models are still based on HMMs. The decoding scheme is more flexible to allow the most common substitutions, deletions and insertions of phones in the language, as described by a phone confusion matrix. This confusion matrix was obtained by comparing the output of the recognizer with the transcription on a larger corpus. The final result for the detection of all disfluencies (word repetitions, stuttering, skipping and mispronunciations) was 44% miss rate at a 5% false alarm rate.

Li et al. (2007) aimed to track children's reading of short stories for a reading tutor. As a language model, they employed a word level context-free grammar for sentences to allow some freedom on decoding words. Each word also had a concurrent garbage model with the most common 1600 words, which aims to detect word level miscues, but was also able to detect some sub-word level miscues for short words. On a detection task of all reading miscues (including breaths and pauses), they achieved a miss rate 23.07% at a false alarm rate of 15.15%.

It should be mentioned that much of the prior research focuses on individual word reading tasks – exceptions being Li et al. (2007) and parts of Yilmaz et al. (2014) –, whereas the present work targets the reading of sentences and pseudowords. As mentioned, some studies go further and attempt to provide an overall reading ability index that should be well correlated with the opinion of expert evaluators (Black et al., 2011; Duchateau et al., 2007), which is also the ultimate objective of our work. These studies always focus on individual word reading tasks, and mainly use reading speed and number of correctly read words to estimate the overall score. Using and analyzing sentences and pseudowords for overall performance scoring is our main contribution and it is expected that, by working with sentences as well as pseudowords, a better understanding of a child's reading ability can be achieved. We also employ new methods to automatically detect disfluencies and explored feature selection and regression models to provide performance scores based on multiple sources of information that can be the ones that teachers consider to evaluate children.

Automatically providing an overall reading aloud performance score for children aged 6–10 years attending primary school is the main objective of this work. For that purpose, a European Portuguese (EP) database of sentence and pseudoword reading recordings was collected and several types of disfluency events were identified. Methods based on task-specific lattices and phone posterior probabilities were developed to annotate data automatically and detect the most common types of disfluencies. Specifically, results on detecting false starts, repetitions and mispronunciations are analyzed. Several features that may be relevant for evaluating performance can be extracted by automatic methods and combined into an overall score. We gathered the opinion of primary school teachers as ground truth for overall performance scores and applied regression models to the extracted features to closely match evaluator opinions. An analysis and selection of features is performed as some features prove to be more relevant than others.

This article is divided into three main sections that are also the key steps necessary for an automatic evaluation of reading aloud, as mentioned above. First, the design and analysis of a database of utterances read by children is described (Section 2), as the type of data used and the disfluencies found are of the utmost importance

for the rest of the study. Next, the automatic speech processing to segment and annotate utterances while detecting several types of disfluencies is described (Section 3). Finally, the procedure to obtain an overall reading performance score is detailed (Section 4): obtaining the ground truth from primary school teachers and using manually and automatically extracted features as input for models that predict reading performance scores.

2. The database

We found it necessary to create a large new speech corpus of European Portuguese (EP) children's speech with utterances of reading tasks that are rich in the common disfluencies that children exhibit while reading. There are some children's speech databases for EP, such as Speecon with rich sentences (Speecon Consortium, 2005); ChildCAST (Lopes, 2012; Lopes et al., 2012) with picture naming; the Contents for Next Generation (CNG) Corpus targeting interactive games (Hämäläinen et al., 2013), and Santos (2014) and Santos et al. (2014) with child-adult interaction. However, these databases do not present the required samples of disfluent read speech. As a first step to collect our data, we undertook a careful design of the presented reading tasks.

2.1. Design

The Portuguese government has defined a set of Curricular Goals (CG) with qualitative and quantitative objectives per grade for reading aloud (Buescu et al., 2015). Some of these objectives include target reading speed of words per minute on short texts, individual words and pseudowords reading tasks. With the analysis of curricular goals in mind, utterances consisting of read sentences and pseudowords were the goal of our data collection. We decided not to include reading of isolated words, as the required time for a session could become too long and the child's performance is likely to decrease with extended sessions. However, a pseudoword reading task was included as it may provide a different and objective analysis of phonetic awareness and letter-to-sound rules independent of word familiarity and context. With sentences, plenty of reading disfluencies can be collected from which the overall reading performance of a child can be evaluated. Each child was presented with a reading task that asked them to read aloud twenty sentences and ten pseudowords. Forty reading tasks were established (10 per grade) to balance repetition and diversity of the data. At a later stage of data collection, these were shortened to 5 tasks per grade, to reinforce repetition. The vocabulary of the set of sentences and pseudowords comprises a total of 2721 word types. The distribution of the material for the different grades is described below.

2.1.1. Sentences

A large set of sentences was extracted from children's tales and school books of the level of the target group (6–10 years old, 1st-4th grades). Selected sentences were mostly short, with a maximum length of 30 words and a mean and standard deviation of 11.1 ± 5.8. Twenty sentences were included in each reading task (for a recording session with one child). The first concern for distributing sentences along the grades was to maintain a good representation of all phones close to their frequency in EP, so that acoustic models of good quality could be built from the data. The other main concerns in building appropriate reading tasks were to maintain the same average difficulty within a grade (with a rising average difficulty from 1st to 4th grades) and to have sentences of varying difficulty within a task (resulting in overlapping distributions of difficulty for different grades). Furthermore, it is necessary to elicit all types of reading disfluencies as training samples, so the difficulty cannot be too low, although a balance must exist so as not to make the tasks unduly hard.

2.1.2. Difficulty

A parameter of difficulty was developed to classify sentences according to phonological and phonotactic constraints. Although it would be ideal to also relate a word's difficulty to its age-of-acquisition or familiarity, not all words of the proposed reading tasks were present in available lexical databases such as ESCOLEX (Soares et al., 2014), and it was not possible to consider such features. The proposed parameter of difficulty is based on the method described in Mendonça et al., (2014), where sentences are evaluated in terms of phonological complexity and variety. All words were split into syllables and a difficulty level was assigned to each syllable, determined by rules based on: the length of the syllable; the multiple pronunciation of some graphemes (e.g. (e.g. ⟨mãe⟩ [mɐ̃j̃] and ⟨bem⟩ [bɐ̃j̃]); the ambiguous pronunciation of consonant clusters (e.g. ⟨prever⟩ [prɘˈveɾ] or ⟨florescer⟩ [fluɾɘʃsˈeɾ]) and vocalic encounters (⟨candeeiro⟩ [kɐd̃iˈɐjɾu] or ⟨veem⟩ [vˈeɐ̃j̃]). Since each syllable has a given minimum difficulty, the length of the sentence also contributes to difficulty.

2.1.3. Pseudowords

Pseudowords (such as ⟨traba⟩ [tɾˈabɐ], ⟨impemba⟩ [iɐ̃ˈeb̃ɐ] or ⟨culenes⟩ [kulˈɛnɘʃ]) represent non-existing or nonsense words which can be used to evaluate morphological and phonemic awareness. A novel method for the creation of pseudowords was developed. Existing tools such as Wuggy (Keuleers and Brysbaert, 2010) take as input existing words and output pseudowords that differ in one or two syllables to the original words. This creates pronounceable words that are similar to existing words (such as ⟨sapado⟩ from <sapato>). The proposed method creates pseudowords without the starting point of valid words while maintaining full pronounceability. It should create non-existing words and the difficulty of reading them should be slightly higher than familiar words. The aim was to create pseudowords of two, three and four syllables. First, the most frequent syllables in each position for words with those number of syllables were extracted from a large lexicon of European Portuguese, CETEMPúblico (Rocha and Santos, 2000). Then, words of two or more syllables are created randomly from a set of the most frequent syllables. Words that have syllabic combinations that do not respect pronounceability rules are deleted, as are words that exist in the lexicon. The difficulty score for a pseudoword is calculated by the same method described above for sentences. The distribution of the pseudowords along the reading tasks is also similar to sentences, promoting a range of difficulty and rising average difficulty along the grades.

2.2. Collection

The corpus of children reading aloud was collected at two private and nine public schools in urban centers and peripheral areas of Portugal's central region with children attending primary school, aged 6 to 10 years. A specific application was developed in which the sentences are displayed in a large font size on a computer screen simultaneously with the start of recording. This presentation allows no practice time that would influence performance. A screenshot of the application can be seen in Fig. 1 as well as an example of the recording environment. The recordings were performed in school classrooms chosen for their low reverberation and noise acoustics. The children were asked to read aloud a set of 20 sentences and 10 individual pseudowords. A lapel Lavalier microphone (Shure WL93) was used as the main recording device, accompanied by a standard table-top PC microphone as backup (Plantronics Audio 10). The background noise could not always be

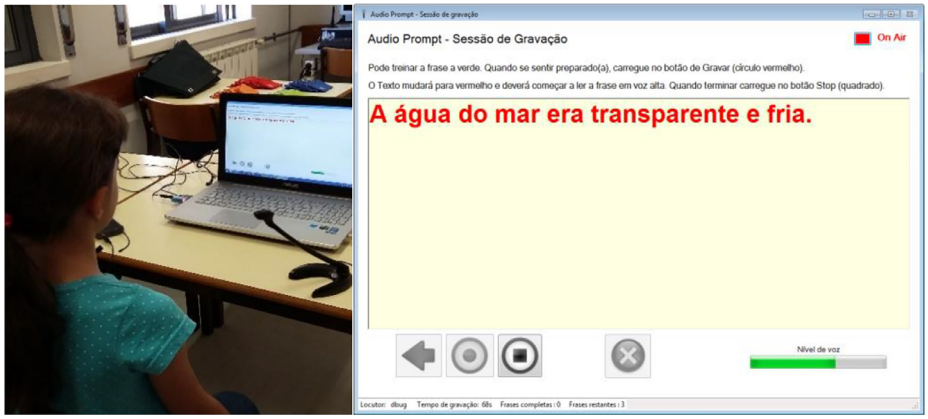

Fig. 1. Example of the recording environment (left) and software (right).

controlled completely, but was mostly low, with an average signal-to-noise ratio of 30 dB, also because the main recording microphone did a good job of filtering out background noise.

2.3. Analysis

The collected database consists of about 20 hours and 7418 utterances of recorded speech from 284 children, 147 female and 137 male, distributed from the 1st to the 4th grade with 68, 88, 76 and 52 children, respectively. A set of 2288 utterances of pseudowords and sentences from 104 children has been fully annotated (5h30m), and these children (46 male and 58 female) are uniformly distributed among the four grade levels (26 per grade). This set is analyzed below in terms of disfluencies and reading speed, and was used as a training set for acoustic models and disfluency detection. A partial annotation of the reading tasks of 75 children was also performed, amounting to 750 utterances (1h31m), used as a test set.

2.3.1. Types and frequency of disfluencies

The annotated speech exhibits a great variety of disfluencies that represent the most common types of errors in reading aloud by children. Based on Candeias et al. (2013), the rules for the annotation and labelling procedure were defined and several types of disfluency were identified as follows:

- PRE – False starts that are followed by the attempted correction (pre-corrections), where multiple can occur. Example: for prompt "*grande espanto*" [gɾˈɐd̃ə iʃpˈɐt̃u], utterance is "*grande* ***espa*** *espanto*" [gɾˈɐd̃ə ˈ**iʃpɐ** iʃˈpɐt̃u].
- SUB – Substitution or severe mispronunciation of a word. Example: for prompt "*voava em largos círculos*" [vuˈavɐ ɐj̃ˈaɾguʃ sˈiɾkuluʃ], utterance is "*voava em* ***lares sicos***" [vuˈavɐ ɐj̃˜**lˈaɾəʃ sˈikuʃ**].
- PHO – Small mispronunciation of a word, usually with a change in one phone or a phone lengthening or extension (EXT, marked with the symbol [:]). Example: for prompt "*A Lena chegou a casa, da escola*" [ɐ lˈenɐ ʃəgˈo ɐ kˈazɐ dɐ iʃkˈɔ.lɐ], utterance is "*A Lena* ***chegou*** *a casa, da* ***escola***" [ɐ lˈenɐ **ʃɔ: gˈo** ɐ kˈazɐ dɐ **ɛʃkˈɔlɐ**].
- REP – Repetition of a word (multiple repetitions may occur). Example: for prompt "*Ele já me deu*" [ˈelə ʒa mə dew], utterance is "*Ele,* ***ele*** *já me deu*" [ˈelə ˈ**elə** ʒa mə dew].
- INS – An inserted word that is not part of the original sentence. Example: for prompt "*mas também dizem*" [mɐʃ tɐb̃ˈɐj̃d̃ˈizɐj̃], utterance is "*mas também* ***me*** *dizem*" [mɐʃ tɐb̃ˈɐj̃˜**mə** dˈizɐj̃].
- DEL – The word was not pronounced (deletion). Example: for prompt "*onde morava* ***uma*** *velha*" [ˈod̃ə muɾˈavɐ ˈ**umɐ** vˈɛʎɐ], utterance is "*onde morava velha*" [ˈod̃ə muɾˈavɐ vˈɛʎɐ].
- CUT – The word is cut off, usually in the initial or final syllable, but not corrected later. Example: for prompt "*dá água ao papagaio*" [da ˈagwɐ aw pɐpɐgˈaju], utterance is "*dá água ao* ***papaga***" [da ˈagwɐ aw **pɐpɐgˈa**].
- PAU (…) – Intra-word pause, when a word is pronounced syllable by syllable with intervening silences. The symbol […] can also appear in other disfluency events denoting a pause. Example: for prompt "*formosa e bonitinha*" [fuɾmˈɔzɐ i bunitˈiɲɐ], utterance is "*formosa e* ***boni…tinha***" [fuɾmˈɔzɐ i **buni…tˈiɲɐ**].

Silence and non-speech events such as breathing, labial and background noise were also annotated. Extensions and intra-word pauses may occur simultaneously with other disfluencies and are marked with [:] and […] in the phonetic transcription. The number of occurrences for each type of disfluency and their percentage of total uttered words in the database are presented in Table 1 for each of the four grade levels.

Some interesting phenomena can be observed, such as 1st-graders being the ones that exhibit more intra-word pauses and extensions (due to slower reading), and 4th-graders having more insertions and deletions (due to faster reading). Furthermore, the defined false start type (PRE) is the most common disfluency for sentences, whereas in pseudowords mispronunciations are more common since there are fewer attempts to correct unknown words. Surprisingly, children did not use filled pauses when trying to read aloud as teen and adults do in spontaneous speech (Veiga et al., 2012), using silent pauses instead when halting their reading.

2.3.2. Reading speed

With annotated data, a simple analysis of the reading performance of each individual child can be done. A common metric is to evaluate reading speed considering only correctly read words, which is defined as Correct Words Per Minute (CWPM) (Hasbrouck and Tindal, 2006). The average values of CWPM per grade of 80 children of our corpus at the end of school year are shown in Table 2, side-by-side with the target curricular goals (Buescu et al., 2015). A large inter-grade overlap of the distributions is observed, showing a variability in reading performance of different children, although the average does increase per grade. Fig. 2 displays this behavior with a boxplot of the distributions of CWPM, showing one clear outlier for the third grade. On data of adult speakers reading (Pellegrini et al., 2013), words per minute average 130.3 ± 17.8. Comparing these values to the observed child performance, there may still be expected improvement from 4th grade children, although some perform as well as adults. For sentence reading, the difference from average CWPM to curricular goals increases in absolute terms along the grades, and these lower CWPM values may be explained by the difficulty of the reading

Table 1
Distribution of disfluency types in sentences for each of the four grades and in pseudowords (number of events and % of total uttered words).

Tags	Sentences					Pseudowords
	1st grade	2nd grade	3rd grade	4th grade	Total	Total
PRE	295 (7.4%)	278 (5.7%)	281 (4.4%)	302 (4.1%)	1156 (5.1%)	318 (15.6%)
SUB	182 (4.6%)	149 (3.1%)	215 (3.4%)	208 (2.8%)	754 (3.3%)	263 (12.9%)
PHO	214 (5.4%)	169 (3.5%)	203 (3.2%)	143 (1.9%)	729 (3.2%)	476 (23.3%)
REP	122 (3.1%)	89 (1.8%)	129 (2.0%)	161 (2.2%)	501 (2.2%)	4 (0.2%)
INS	30 (0.8%)	42 (0.9%)	42 (0.7%)	65 (0.88%)	179 (0.8%)	20 (1.0%)
DEL	5 (0.1%)	14 (0.3%)	16 (0.3%)	50 (0.68%)	85 (0.4%)	3 (0.2%)
CUT	11 (0.3%)	15 (0.3%)	29 (0.5%)	27 (0.37%)	82 (0.4%)	2 (0.1%)
EXT:	256 (6.5%)	145 (3.0%)	212 (3.3%)	73 (1.0%)	686 (3.0%)	431 (22.7%)
PAU…	179 (4.5%)	126 (2.6%)	102 (1.6%)	65 (0.9%)	472 (2.1%)	251 (13.1%)

Table 2
Per grade mean and standard deviation of measured Correct Words per Minute (CWPM), Curricular Goals (CG) of CWPM and relative difference of CWPM to CG, for sentences and pseudowords reading tasks.

Grade	Words in sentences			Pseudowords		
	CWPM	CG	CWPM-CG	CWPM	CG	CWPM-CG
1st	59.7 ± 18.1	55	+8.5%	18.8 ± 8.0	25	−24.8%
2nd	85.2 ± 22.9	90	−5.3%	26.7 ± 8.4	35	−23.7%
3rd	97.1 ± 23.5	110	−11.7%	26.1 ± 6.5	–	
4th	110.4 ± 22.7	125	−16.7%	34.9 ± 9.6	–	

Table 3
Mean and standard deviation per grade of pseudoword reading reaction times (in seconds), number of uttered words with any kind of disfluency event (including extensions and intra-word pauses) and number of incorrect words.

	1st grade	2nd grade	3rd grade	4th grade
Reaction Ttime (s)	1.65 ± 0.83	1.35 ± 0.43	1.14 ± 0.23	1.19 ± 0.35
Number of disfluent words (out of 10)	6.54 ± 2.89	3.23 ± 2.32	2.96 ± 1.87	2.70 ± 2.24
Number of incorrect words (out of 10)	4.29 ± 2.33	2.31 ± 2.06	2.19 ± 1.57	2.17 ± 1.92

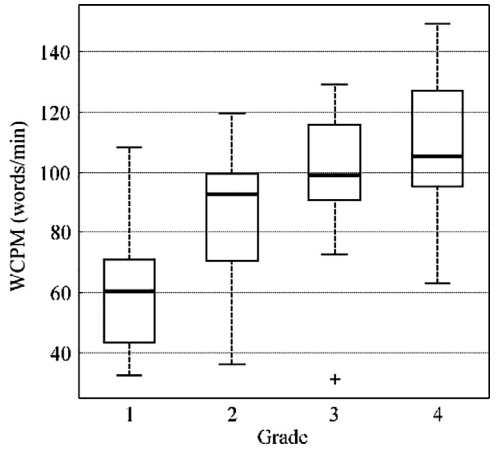

Fig. 2. Median and quartiles boxplots of Correct Words per Minute (CWPM) for sentence reading tasks for each of the four grade levels.

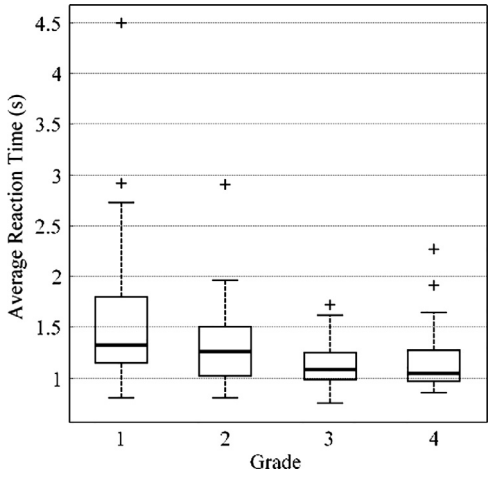

Fig. 3. Median and quartiles boxplots of average Reaction Times for the pseudoword reading task for each of the four grades.

tasks. It can be concluded that the suggested increase of difficulty along the grades may be too steep to directly evaluate CG as intended, and, for overall reading ability evaluation, this difficulty needs to be taken into account. For pseudowords, although there are no CGs for the third and fourth grades, average CWPM values are significantly lower than the CG, suggesting that the generated pseudowords (by joining common syllables and not based on existing words) are of high difficulty.

The defined curricular goals can be a starting point to appraise a child's reading ability. However, they do not take into account factors such as task difficulty or type of disfluencies; therefore, other ways to qualify reading performance should be considered.

2.3.3. Pseudoword performance and reaction time

To further analyze children's performance on the task of reading individual pseudowords, data from 100 children is considered, in which they read 10 individual pseudowords, one at a time per recording. This task differs substantially from sentence reading as morphological and phonemic awareness are the factors that influence a good performance on reading unknown words. Several interesting metrics can be extracted here, which may contribute to overall reading performance. First, the reaction time of starting to read the word (the time between the start of presentation and the onset of speech) reflects how fast the child achieves confidence in reading the entire word or the first syllable, especially for first graders. However, this metric does not reflect whether the word is read correctly or not, and there are children with fast reaction times who do make several mistakes. Still, the average reaction time decreases along the grades, as observed in Table 3 and Fig. 3, with only a small increase from third to fourth grades.

Also in Table 3, the number of words that had any disfluency event is listed. For the first grade, the average of 6.5 disfluent words out of 10 is much higher than for other grades. Note that this measure is not identical to number of incorrect words (also presented in Table 3), since phone extensions or intra-word pauses may occur.

3. Automatic annotation and detection of disfluencies

The challenge of automatically processing utterances of read child speech was approached as a two-step process. First, an alignment (or segmentation) that considers the original prompt and allows extra content based on the original words is applied. With the resulting segments, a classification stage determines if a word was correctly pronounced or not. The annotated set of 2288 utterances from 104 children amounting to 5h30m of audio was used in this section as a training set. The test set corresponds to 1h31m from 750 utterances of 75 children.

3.1. Alignment

False starts and repetitions represent most of the occurring extra segments in our annotated data: 91% of extra-events (false starts, repetitions and insertions), as can be computed from Table 1. As such, for the automatic annotation, we decided to apply a first stage to align the data as best as possible to word-relevant segments. One problem is not considering mispronunciations, but it is still hoped that by forcing the original word to be aligned to the mispronounced segment, correct time stamps may be obtained. For this stage, the Kaldi system (Povey et al., 2011) was used both to train acoustic models using only the manually annotated train set and to perform the decoding. From previous work (Proença et al., 2015), it was found that using a small amount of child speech data to train acoustic models was better than using models trained with a large amount of adult speech adapted to child speech. We used standard triphone models with 12,000 Gaussians.[1]

The proposed method consists of the following steps:

1. Voice activity detection is applied to the audio to deal with intra-word pauses, and pauses longer than a given threshold are removed.
2. A specific word-level lattice for each given sentence or task is built.
3. Decoding is performed using the specific lattices, obtaining the best label/segment sequence.
4. A reconstruction of the alignment is done, taking into account the silences previously removed.

Intra-word pauses occur when words are pronounced syllable by syllable with intervening silences, most often for first grade children. It is hard for the decoder to align a word when silence exists between syllables. Thus, we apply a voice activity detection method to cut silent segments. Even if silence between words is cut, which would help to clearly separate them, results improve due to the amount of intra-word pause cases solved. We detect the silence segments by analyzing the smoothed logarithmic energy of the signal, and selecting low energy segments longer than 150 ms, using a moving threshold that is a function on the high and low energy levels in the signal.

For the decoding stage, we build task-specific lattices to allow some of the common patterns found in the data when considering false starts and repetitions. Finite state transducers (FST) are used and, compared to previous work (Proença et al., 2015), there are fewer possibilities for word sequence repetition, avoiding any back-transitions that complicate the FST and increase decoding times, which would be undesirable for a live application. As an example, Fig. 4 describes the FST grammar for the three-word sentence *ele sonhava muito* [ˈelə suɲˈavɐ mˈujt̃u] (“he dreamed a lot”). There is a basic group of FST nodes representing each word, allowing false starts and repetitions of the word. For example, the word “ele” is represented by the nodes 0, 1 and 2, and the transitions arriving at nodes 1 or 2. The same applies to the word “sonhava” with nodes 2, 3 and 4, and for the word “muito” with nodes 6, 7 and 8.

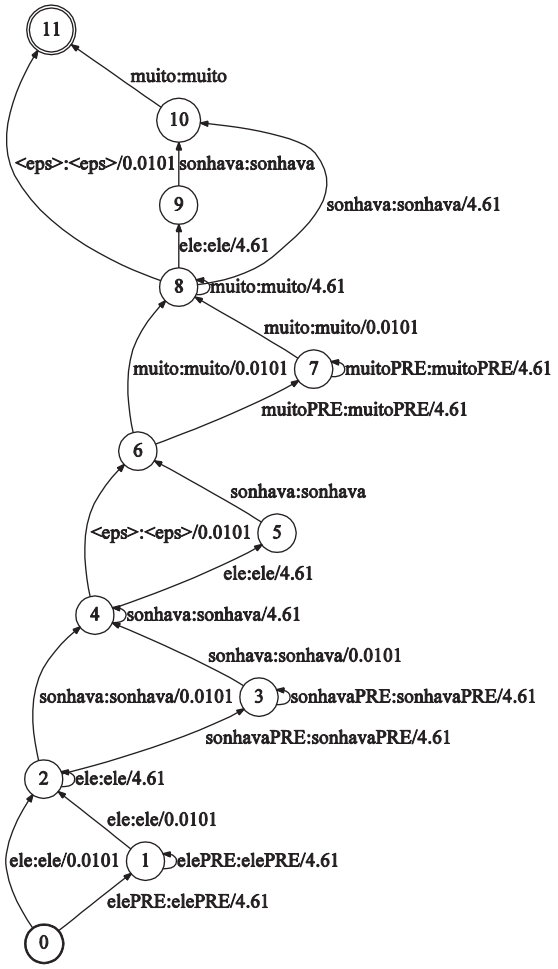

Fig. 4. Sentence FST for the three-word sentence *ele sonhava muito* [ˈelə suɲˈavɐ mˈujt̃u].

The word units with the suffix “PRE” represent syllable-based false starts (pre-corrections). Although many false starts are mispronounced, they often correspond to interrupting the pronunciation attempt at syllable boundary. Therefore, these cases are considered with multi-pronunciations for PRE up to (and excluding) the last syllable, e.g., for a four-syllable word: the first syllable; the first followed by the second; or the first, second and third consecutively. Specifically, *elePRE* can only be [e]; *sonhavaPRE* can be [su] or [suɲˈa]; *muitoPRE* can only be [mujt̃].

In addition, we included the possibility of repeating the previous sequence of two or three words, at most. This kind of occurrence, e.g., *ele sonhava ele sonhava muito* is very common in the data, and often represents an attempt to correct a mistake by restarting at a sentence or clause boundary. In the example FST, paths that go through nodes 5, 9 and 10, represent these possibilities. Furthermore, following the left-most arcs, one gets the original sentence without any false starts or repetitions (<eps> is an epsilon arc, consuming no input or output).

Other than not accounting for mispronunciations, a limitation of the described method is not allowing for deletions or insertions. In fact, these are not very common in the data, as children practically always try to finish reading the sentence. For a more general application, it may be preferable to allow deletions (skipped

[1] Acoustic models with neural network alternatives, trained with Kaldi, did not improve results for this task, probably due to the relatively small amount of training data.

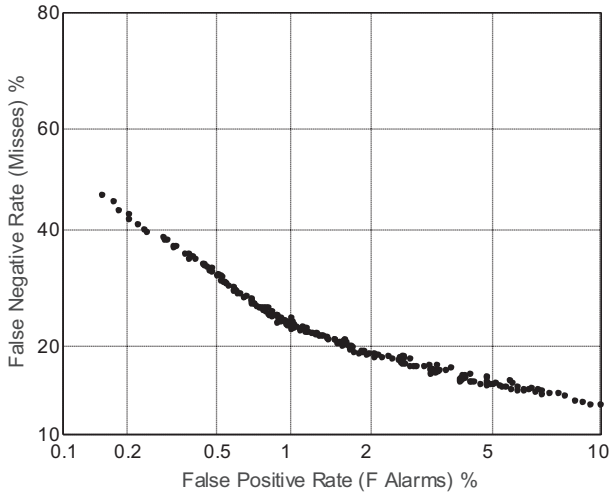

Fig. 5. Detection error tradeoff (DET) curve for the detection of false starts and repetition events on the training set.

words), accounting for cases where a sentence is only partially pronounced.

After the segmentation and forced alignment, we need to classify each segment as correctly pronounced or not.

3.1.1. Results

The WER using the text of the original prompts as hypothesis and the segmentation of the manual transcription as reference is only 9%. This means that repetitions, false starts, insertions and deletions, occurring in the manual transcription, account for these 9%. By using the described method that allows repetitions and false starts to be found, the best WER achieved in the training set was 3.75%. To evaluate the system's performance in detecting events (PRE or REP) in terms of misses (false negatives) and false alarms (false positives), we consider that:

- Extra detected events are false alarms;
- Any undetected event is a miss;
- An event erroneously detected as an event of an adjacent word (a substitution) is also a miss.

These stipulations are similar to those used in NIST evaluations (Fiscus et al., 2007), although to calculate the false alarm rate we divide the number of false alarms by the number of original words. By using a wide search beam during decoding and varying the word insertion penalty and lattice rescoring weights, a Detection Error Tradeoff (DET) curve can be obtained, as presented in Fig. 5, for the training set.

The best WER obtained (3.75%), corresponds to a very low false alarm rate of 0.89% and a 23.53% miss rate. Comparing to a previous method (Proença et al., 2015), where 30.62% miss rate is obtained for the same false alarm rate, this represents a 23% relative improvement in miss rate for this operating point. Using the word insertion penalty and rescoring weight from the best WER, the results on the test set are: 4.47% WER, 1.94% false alarm rate and 20.60% miss rate. The best possible WER would be 4.01%, by choosing an optimal word insertion penalty. Certain aspects of our system can account for errors in specific event labeling: for small words with one syllable only repetitions are marked; and since some PRE tags of larger words are mispronunciations of the whole word, they can be decoded as the word followed by repetitions. Furthermore, there are insertions that are never accounted for, always leading to false alarms or segment mismatches, and lowering overall accuracy.

This stage outputs a time-stamped alignment of the data according to word-relevant segments, which serves as input to the next classification stage.

3.2. Mispronunciation detection

In order to detect mispronunciations we trained a neural network for phoneme recognition, using the Brno University of Technology neural network system,[2] which is based on long temporal context. The manually annotated training set was used for training the neural network, achieving about 70% phoneme recognition accuracy on the test set. With this neural network we obtained the posterior probabilities of the phoneme model states for all sentences of the database, the so-called posteriorgrams. These posteriorgrams could be used as input for the FST decoder of Section 3.1. However, doing so did not improve the results compared to using Gaussian mixture models. For mispronunciation detection, the trained neural network provided better results due to better posterior probabilities. These results were also better compared to using neural network models trained with Kaldi.

Using the posteriorgrams, we can use a word spotting system to try to detect correctly pronounced words. The word spotting system is based on the log likelihood ratio (LLR) between the spotting model (the sequence of phonemes of the spotting word) and a filler model that consists of a loop of all phoneme models (Veiga et al., 2014). The token-passing paradigm is used to compute the likelihoods and track the starting time of the tokens at the output of the word spotting model. A match is detected if a peak value of the LLR is above a given threshold.

We apply the word spotting system to each word hypothesis given by the previous alignment, and find the peak LLR in the close vicinity of the given alignment. Several intervals to define this close vicinity were tested. If the peak LLR of a word is below a certain threshold, it is classified as mispronounced. The trade-off of false alarm rate versus miss rate on mispronunciation detection can be represented with a Detection Error Tradeoff (DET) curve by varying the decision threshold, as described below.

3.2.1. Results

During manual annotation, two classes of mispronunciations of different severity were considered: PHO – variations of only one phoneme; and SUB – severe mispronunciation or substitution of the word. For the mispronunciation decision task, we present results using two ground truths: SUB or PHO segments as a mispronounced class (SUB+PHO) versus correctly pronounced words; and only SUB versus correct words (since PHO is usually too difficult to detect). We consider both the manual and the automatic segmentation to define segments for mispronunciation classification. In the case of automatic segmentation, we allow some misalignments with the ground truth. However, segments must overlap in order to be considered as matches to a particular ground truth segment.

The discriminant to decide mispronunciation is the maximum LLR of word spotting in a segment and we considered several intervals around the final time of an aligned segment to search for the maximum LLR. Given that there are some misaligned segments, this proved to be a better approach than calculating LLR using the time stamps of the segmentation. An optimization revealed that the best interval for using manually annotated segments was -100 ms to $+50$ ms and for the automatic segments -250 ms to $+50$ ms. We also experimented with several LLR score normalizations: dividing by the number of phones of the searched

[2] Phoneme recognizer based on long temporal context, Brno University of Technology, FIT. http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

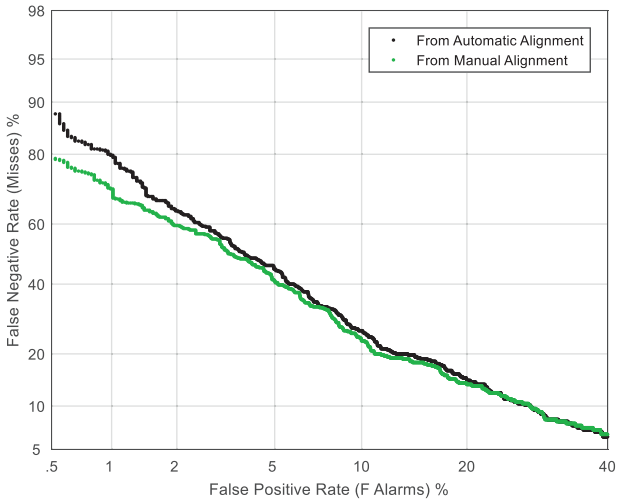

Fig. 6. Detection error tradeoff (DET) curves for mispronunciation classification of SUB+PHO class on the test set.

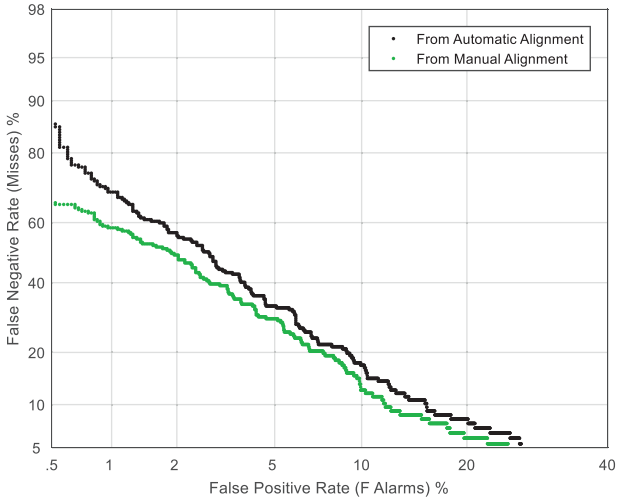

Fig. 7. Detection error tradeoff (DET) curve for mispronunciation classification of SUB class on the test set.

word; dividing by the number of frames occupied by the best spot; and dividing by the LLR area of the spot as described in Veiga et al. (2014). All of these normalized scores benefited from adding an extra value: the original LLR score scaled with a small constant factor (the optimal factor varying per normalization approach). By doing this, the results are very similar, with normalization by the number of phones having a slight advantage. This is the one used in the results presented below.

Fig. 6 presents the DET curve for the SUB+PHO vs. correct words classification, using manual or automatic segmentation. Fig. 7 presents the results for the SUB vs. correct words classification. For a false alarm rate of 5%, miss rates of 40.99% and 44.94% are obtained for the SUB+PHO class, and 28.88% and 32.62% for the SUB class, for manual and automatic alignments, respectively. We target a low false alarm rate that still provides miss rates under 50%. The goal was to be lenient with the child reader, allowing some non-detections instead of generating frequent false alarms.

As expected, miss rates for using the automatic alignment are slightly worse, although still very close to manual alignment, as reflected in the closeness of the DET curves. Nevertheless, we believe that these results could be improved by using a fusion of different scores or different normalizations.

After applying the proposed methods, we obtain both an automatic detection of the number of disfluencies per utterance and an automatic annotation with the suggestion of phone sequence for mispronounced segments. Most metrics that could be obtained manually can also be extracted from the automatic output, such as correct words per minute.

4. Overall reading performance score

Although measuring correct words per minute can already be one way to evaluate a child's reading, there may be other factors or specific problems that characterize the child's performance. Computing a score based on features from sentence reading tasks and pseudoword reading tasks can hopefully give an improved overall assessment of reading performance. We have gathered the opinion of primary school teachers as ground truth for overall performance and built regression models based on several features extracted from child utterances, while comparing the use of manual and automatic annotations.

4.1. Ground truth

In order to obtain a professional assessment for reading ability in children, we asked primary school teachers, through a targeted crowdsourcing effort, to listen to utterances of reading tasks of children and provide a score for overall performance. These opinions will be used as a ground truth with which our computed scores should be well correlated. A total of 150 children from the collected dataset were evaluated, 43 from the first grade, 40 from the second grade, 35 from the third grade and 32 from the fourth grade.

We aimed that each evaluator should not spend more than 30 minutes on the requested task and it was necessary to balance this time limit with how many raters a child could be evaluated by. In our corpus each child reads 20 sentences and 10 pseudowords. Initial tests for the rater effort showed that listening to only 5 sentences and 5 pseudowords is enough to provide an overall performance score, with the score rarely changing if more utterances are listened to. This assumption allowed us to realistically aim for each child to be evaluated by at least 5 teachers. In the end, an average of 10 teachers evaluated each child.

We have 10 groups with a variable number of evaluators (7 minimum, 13 maximum, 10 average), for a total of 100 evaluators. The number varies as our collection process was affected by some allocated evaluators not finishing evaluations (not counted here). Each group evaluates the same set of 15 children (for a total of 150 children) and each set is different for each group (but well balanced for grade levels). Therefore, although results may be shown for all evaluators, the calculations must be done separately for each group.

4.1.1. Evaluating evaluators

For each evaluator, we can compute Pearson's correlation of the 15 scores given to the 15 scores of another evaluator who gave scores for the same children, repeating for all evaluators of the same group. This measure reflects pairwise agreement between evaluators. For a group of 10 evaluators, there would be 9 values per evaluator. The mean of these 9 values for each evaluator describes their overall agreement with the group, shown in Fig. 8.

There is one clear outlier with an average correlation of 0.413; this evaluator should be removed. It can be argued that the next 6 with low correlations also stand out, and the evaluators below 0.65 correlation should probably be removed for computing the ground truth. A problem is that some of them belong to the same group and by removing the worst of them, the others' average can

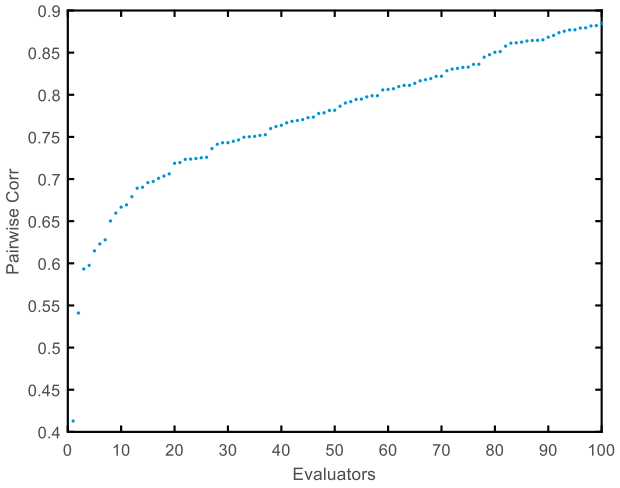

Fig. 8. Mean pairwise correlations by evaluator, sorted from lowest to highest.

Table 4
Final overall mean and standard deviation values of pairwise correlation and correlation to the mean of other evaluators for 100 evaluators.

Correlation	Mean ± S.D.	Maximum	Minimum
Pairwise	0.796 ± 0.060	0.885	0.657
To the mean of others	0.874 ± 0.069	0.967	0.679

improve and be higher than the threshold. Additionally, the pairwise correlations of all other evaluators must also be computed again. By removing the worst evaluators iteratively until none below 0.650 are kept, only 5 are eliminated.

As an alternative to pairwise correlation, the correlation of an evaluator's 15 scores with the 15 mean scores averaged from the other evaluators of the same group may be used. It gives higher values than the pairwise correlation but conclusions are similar. The final values for the two metrics are shown in Table 4.

4.1.2. Normalizing scores

As an alternative to using a mean score for a child from the raw values given by teachers, applying a z-normalization (z-norm) per evaluator, as in (1), can remove certain biases. These effects for an evaluator can be: i) constantly giving lower scores than the average ones; ii) constantly giving higher scores than the average ones; iii) constantly giving scores near the minimum and maximum; or iv) constantly giving middling scores.

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

The z-norm for each evaluator changes their scores (x) by subtracting the mean of their 15 scores (μ) and dividing by the standard deviation of the 15 scores (σ). This gives values with zero mean and unitary standard deviation. Since these values do not fall in the intended scale of 0–5, they need to be reconstructed. We do this by multiplying by the overall standard deviation of all scores (1500) and adding the overall mean. This method can provide values slightly lower than 0 or higher than 5 (an alternative would be to scale the minimum to 0 and the maximum to 5).

A pairwise correlation analysis would provide similar results to using non-normalized scores, since the changes are linear and correlation is linear. The same evaluators are removed. Fig. 9 compares the final scores obtained using z-norm to the ones from a simple mean of raw scores. The standard deviation of a child's scores (the 7 or more scores given by teachers) also lowers from an average 0.719 to 0.549 with normalization.

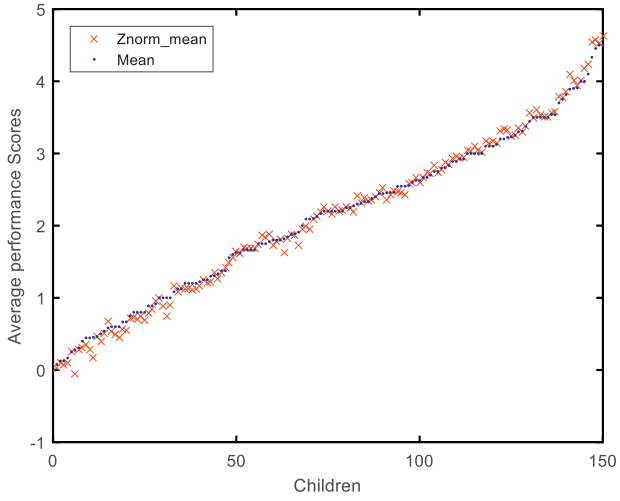

Fig. 9. Mean of raw evaluator scores vs. mean of z-normalized scores. Mean differences are 0.062 ± 0.057, with a maximum of 0.329 (a score of 0.273 becoming −0.057).

Table 5
Enumeration of features. These are extracted separately for sentence tasks and pseudoword tasks, leading to 52 features. Those with an asterisk are not computed for the automatic methods.

Feat #	Abbreviation	Description
1	WPM	Words per minute (original prompt)
2	CWPM	Correct Words Per Minute
3	SyllsPM	Syllables per Minute (original prompt)
4	CSPM	Correct Syllables per Minute
5	CharsPM	Characters per Minute (original prompt)
6	CCPM	Correct Characters per Minute
7	SILrate	Rate of Silence (Total Silence / Total time)
8	SILini	Average Initial Silence time before first word
9	SILiniRate	Initial Silence time / Total Time
10	SUBrate	Rate of SUB (number of SUB events / number of Words)
11	PHOrate	Rate of PHO (number of PHO events / number of Words)
12	PRErate	Rate of PRE (number of PRE events / number of Words)
13	REPrate	Rate of REP (number of REP events / number of Words)
14	PAUrate	Rate of PAU (number of PAU events / number of Words)
15	DELrate *	Rate of DEL (number of DEL events / number of Words)
16	EXTrate	Rate of EXT (number of EXT events / number of Words)
17	INSrate *	Rate of INS (number of INS events / number of Words)
18	MispR	Rate of SUB+PHO (Mispronunciations)
19	ExtraR	Rate of PRE+REP (Extra segment disfluencies)
20	SlowR	Rate of PAU+EXT (Slow reading disfluencies)
21	FastR *	Rate of DEL+INS (Fast reading disfluencies)
22	DisfR	Rate of Disfluencies (sum of all events / number of Words)
23	nSylls	Total number of syllables (original prompts)
24	nChars	Total number of characters (original prompts)
25	Diff1	Difficulty 1 – Pronunciation rules without counting length
26	Diff2	Difficulty 2 – Original difficulty index (rules and length)

4.2. Features

In an attempt to explain which characteristics teachers take into account when deciding on an overall performance score, we will analyze how well certain features fit to the ground truth score and try to get the best possible fit with a combination of these features. Two separate analyses are done, using two sets of features: one set extracted from manual annotations and the other from the automatic annotation described in the previous section. Although features from manual annotation may give the purest conclusions on what is indeed significant for reading performance, the automatically obtained features are the ones that will prove if we can assess performance without human intervention.

The full set of considered features is described in Table 5. The same features are extracted from sentence reading tasks and pseu-

doword reading tasks separately, doubling the number of features shown. Features can be split into four groups: reading speed 1–6, silence related features 7–9, rate of disfluencies 10–22, and task-specific information 23–26. Since there are a couple disfluency types that were not targeted in the automatic methods (deletions and insertions), features that depend on these disfluencies are only computed and analyzed for the manual annotation. From this point onwards, since features 1–26 repeat for sentences and pseudowords, we will address sentence features by prefixing an 's' and pseudoword ones with a 'p' (e.g., s1, s2, p1, p2).

An additional feature – Grade – that describes the school grade level each child is enrolled in (1 to 4) will also be analyzed. However, since a final application may not want to require knowledge of a child's grade level, we should build models that do not use this feature. Evaluators did not have grade information, although it is foreseeable that it will have some correlation to given scores, as average performance increases per grade (as previously observed for reading speed).

Some of the considered features are clearly very similar to each other, and some interesting conclusions may be drawn from analyzing their pairwise linear correlations. Given that this analysis results in a 53 by 53 symmetric matrix, these are the main observations for manually obtained features:

- Features s1-6 (reading speed) are highly correlated between each other (>0.95).
- From p1-6, p5 is the most correlated with s1-6 (always ≈ 0.8), which is also the highest correlation between sentence and pseudoword features.
- Rate of silence in sentences (s7) is inversely correlated to reading speed s1-6 (≈ -0.69).
- Rate of disfluencies is significantly correlated with non-disfluency-related reading speed: s1 and s22 with -0.71; p1 and p22 with -0.61.
- As expected, Grade is significantly correlated with task information s23-26 ($0.80 < \rho < 0.84$) and with reading speed: s1-6 ($0.67 < \rho < 0.72$); p1-6 ($0.45 < \rho < 0.52$).

Since some of these features provide similar information, it is foreseeable that between highly correlated features such as s1-6, one of them will be sufficient to predict reading performance. The next step in feature analysis will be to determine how each of them individually is able to predict ground truth scores.

4.3. Individual feature performance

The simplest way of fitting a feature to ground truth scores is by applying a linear transformation as in (2), trained by a linear regression (LR) model that minimizes the sum of squared errors (least squares).

$$\hat{y} = a^T X + b \tag{2}$$

In Eq. (2), $\hat{y}$ is the predicted output, X is the feature matrix (each column is a feature vector), a is the coefficient (weight) of the input feature and b is the intercept (bias) term. Two metrics that evaluate the fit of a model to the ground truth will be considered: Pearson's correlation coefficient (ρ or Corr) and root mean squared error (RMSE) as described by Eqs. (3) and (4). In both equations, $\hat{y}_i$ is the predicted output for a child, y_i is the reference score (ground truth given by the mean of normalized scores of teachers), $\mu_{\hat{y}}$ and μ_y are the mean scores for predicted outputs and ground truth and n is the number of children/scores analyzed ($n=150$ in our case).

$$\rho = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \mu_{\hat{y}}\right)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n} \left(\hat{y}_i - \mu_{\hat{y}}\right)^2} \sqrt{\sum_{i=1}^{n} (y_i - \mu_y)^2}} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{n}} \tag{4}$$

To train and test regression models, we will consider a leave-one-out cross-validation with 150 folds, where 149 subjects are used to train a model and 1 is left out for testing, until every subject is used in testing. Corr and RMSE are calculated with 150 resulting test values, gathered from the different folds. Although it is cumbersome to train 150 models, it is the best way to avoid dependence on different randomizations of folds that would lead to different average results.

Table 6 indicates the performance of each feature if used individually to train a linear model. Random performance leads to a correlation coefficient of 0 and RMSE of 1.90. None of the strong correlation values are negative since any negatively correlated features are transformed with the linear model with a negative coefficient a.

The best overall feature for both manual and automatic methods is s6: correct characters per minute in sentences. A correlation of 0.94 for the manual feature indicates that this metric by itself can be a very good predictor of overall reading performance, proving that evaluators focus mostly on reading speed. Features based only on the number of disfluencies over the number of words (such as s22 – DisfR), although presenting a correlation around 0.7 for sentences, do not perform as well as reading speed, which shows that reading speed is of higher importance. Reading speed metrics that do not depend on detecting disfluencies (s1, s3 and s5), although performing very well, are slightly worse than their counterparts that depend on disfluencies. For pseudowords, the opposite occurs, with the non-disfluency-dependent reading speed features having significantly better performance. This may be due to some very poor performances in the pseudoword task, where the time it took to read them conveys more information than the number of correct readings (values of 0 correct words per minute can be found). Fig. 10 shows the relation between ground truth scores and the best features for sentences and pseudoword tasks, for the manual case, where there is evidence of a linear fit, especially for s6.

It must be emphasized that the results obtained with features 1, 3 and 5 can be dependent on the conditions of our data. In our collected dataset, it is typical that reading tasks are completed even if a lot of mistakes are made. It is very rare that a sentence is not finished and pseudoword lists are always completely attempted, which leads to these reading speed metrics, which only take into account the original prompt without considering if there are disfluencies, to have a certain significance for reading speed. If this were not the case, features 2, 4 and 6 would be clearly preferable, since unfinished attempts or nonsense attempts would severely influence them. In a live application, these types of attempts should be expected. Additionally, there may be other cases where reading speed or correct words/characters per minute are not enough to characterize reading performance. For example, a very fast reader who often repeats words or gives a lot of false alarms but ends up pronouncing words correctly could have the same CWPM value as a reader with normal speed who reads without disfluencies. Even for incomplete attempts, CWPM could be of normal value, since it doesn't take into account the deleted words. For these cases, the features based on the relative number of disfluencies (e.g., s22 and p22) could be of help.

For the automatic features, the same conclusions apply, although s6 performs slightly worse than its manual equivalent. Unexpectedly, the performances of disfluency-dependent reading speed features (s2, s4 and s6) also fall closer to non-disfluency ones (s1, s3 and s5) probably due to certain disfluency detection

Table 6
Performance of linear regression models predicting ground truth scores using individual features (average of leave-one-out cross-validation).

Sentences						Pseudowords				
Feat.	Abbr.	Manual		Auto			Manual		Auto	
		Corr	RMSE	Corr	RMSE	Feat.	Corr	RMSE	Corr	RMSE
s1	WPM	0.919	0.459	0.917	0.463	p1	0.744	0.776	0.744	0.776
s2	CWPM	0.928	0.434	0.923	0.447	p2	0.674	0.858	0.670	0.863
s3	SyllsPM	0.927	0.435	0.926	0.439	p3	0.764	0.750	0.760	0.755
s4	CSPM	0.938	0.402	0.930	0.429	p4	0.684	0.848	0.684	0.848
s5	CharsPM	0.931	0.424	0.930	0.428	p5	**0.805**	**0.689**	**0.803**	**0.693**
s6	CCPM	**0.940**	**0.397**	**0.931**	**0.425**	p6	0.703	0.827	0.691	0.840
s7	SILrate	0.647	0.885	0.736	0.787	p7	0.324	1.099	0.397	1.067
s8	SILini	0.347	1.091	0.480	1.019	p8	-0.157	1.176	-0.130	1.176
s9	SILiniRate	0.283	1.115	0.231	1.132	p9	-0.073	1.173	0.005	1.169
s10	SUBrate	0.376	1.105	0.615	0.916	p10	0.397	1.066	0.494	1.011
s11	PHOrate	0.394	1.073	N/A	N/A	p11	0.031	1.167	N/A	N/A
s12	PRErate	0.547	0.973	0.577	0.951	p12	0.131	1.156	0.217	1.135
s13	REPrate	0.190	1.142	0.378	1.076	p13	-0.008	1.170	0.010	1.170
s14	PAUrate	0.457	1.036	0.328	1.098	p14	0.124	1.156	-0.172	1.192
s15	DELrate	-0.037	1.172	N/A	N/A	p15	0.109	1.164	N/A	N/A
s16	EXTrate	0.350	1.092	0.381	1.073	p16	0.174	1.145	0.189	1.135
s17	INSrate	0.234	1.130	N/A	N/A	p17	-0.051	1.173	N/A	N/A
s18	MispR	0.525	0.991	0.615	0.916	p18	0.389	1.070	0.494	1.011
s19	ExtraR	0.498	1.008	0.555	0.968	p19	0.139	1.154	0.236	1.130
s20	SlowR	0.540	0.978	0.328	1.098	p20	0.247	1.127	-0.172	1.192
s21	FastR	0.171	1.146	N/A	N/A	p21	0.092	1.160	N/A	N/A
s22	DisfR	0.663	0.872	0.683	0.850	p22	0.490	1.013	0530	0985
s23	nSylls	0.456	1.034	0.456	1.034	p23	0.147	1.151	0.519	0.993
s24	nChars	0.483	1.018	0.483	1.018	p24	0.361	1.084	0.147	1.151
s25	Diff1	0.493	1.011	0.493	1.011	p25	0.464	1.029	0.361	1.084
s26	Diff2	0.491	1.012	0.4909	1.0123	p26	0.462	1.031	0.464	1.029

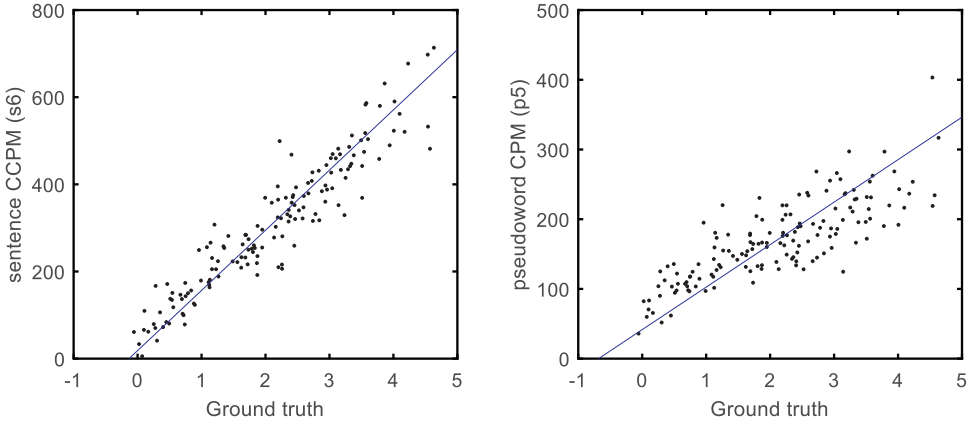

Fig. 10. Ground truth scores vs. the best sentence feature (s6, left) and the best pseudoword feature (p5, right) for manually obtained features, including their linear regression lines.

errors, which may lead to the conclusion that it is not necessary to detect disfluencies. Again, if our data presented nonsense attempts, the results should be different.

Using the Grade feature for a linear regression model gives a positive weight, 0.647 Correlation and 0.886 RMSE, showing that scores increase per grade on average.

Although the best feature – CCPM in sentences – already aggregates a lot of information (total reading time, length of tasks, length of words, correctly pronounced words), we aim to improve models of overall score by using information from additional features in the same model.

4.4. Multi-feature models

To use multiple features in a regression model, we explore linear regression (LR) and Gaussian process regression (GPR) over several feature selection methods. One of the main problems to be tackled is overfitting, since a linear regression considering all the defined features will be strictly optimized for the training sets, with no regards for generalization. The selection of the most relevant features can be a way to minimize overfitting as well as other regression methods such as GPR.

GPR builds kernel-based probabilistic models to infer continuous values and is especially useful to avoid overfitting (Rasmussen and Williams, 2006). Since it is probabilistic, confidence intervals on a provided score can also be calculated. We trained GPR models with a squared exponential kernel as the covariance function.

Stepwise regression can iteratively decide which features to include or remove for a regression model (Draper and Smith, 1998). We explored two stepwise approaches that start with no features included: only adding features (add, forward or sequential) and bidirectional where features can be added and later removed (add+remove). The criterion to add a feature at each step is select-

ing the one that minimizes the sum of squared errors (SSE) when a linear regression is applied. However, a feature is only added if the decrease in SSE is statistically significant with a p-value of an F-statistic test lower than 0.05. Similarly, a feature can be removed at a certain step if its contribution to lowering SSE at this stage is not statistically significant. Either LR or GPR can be applied to the selected features.

Least absolute shrinkage and selection operator (LASSO) is a regularization technique that we apply here for a regularized least-squares regression (Tibshirani, 1996). LASSO minimizes SSE but adds constraints to the sum of absolute values of coefficients of features, usually producing many weight coefficients equal to zero, usually for highly-correlated features. It produces a solution for a linear transformation but can also be a feature selection procedure (by selecting the features with weights different from zero), after which LR or GPR can be applied.

We also explore principal component analysis (PCA) to transform features into a set of linearly independent ones (Jolliffe, 2002). By applying this transformation to the entire set of features, it is hoped that the newly created features, especially the ones that explain most of the variance of the data, can be useful for regression. We apply PCA to the entire set of features, and select the ones that explain 95% of the variance to train LR and GPR. Additionally, we apply bidirectional stepwise regression to the entire set of new features.

Using the stepwise algorithm for different cross-validation folds may result in different selections of features, with some being selected in several or all folds. With this knowledge, we analyzed the results of pre-selecting the most common features given by the stepwise folds and then running LR or GPR using only those features. Although this feature selection depends on all the cross-validation folds using stepwise selection, the actual models trained (with a new leave one out cross-validation) do not depend on the left-out test values. These selections can be different for manual and automatic features. However, for manual features, the stepwise algorithm using leave-one-out folds selected the same four features 100% of the time, unlike with automatic features, where variations did occur. So, for the manual case, the following selected features are from a stepwise algorithm that allows the addition of features with an increased p-value for the F-test (0.15), now with variations occurring. The selections made are of features that were chosen in stepwise manner for at least a certain percentage of the folds: 80% (Sel80%), 40% (Sel40%) and 5% (Sel5%).

Table 7 summarizes the results of all the multi-features models explored when considering manual or automatic features. The first two values are of the best individual feature (s6 – CCPM) and the selection of the best feature of sentences (s6) and the best of pseudowords (p5 – CPM).

The most notable observation is that GPR proved to be superior to LR at every stage, demonstrating its generalization capabilities. Successful feature selections only improved results slightly compared to using GPR over all features, which is already very robust to noise. The improvement of using multiple features instead of only s6 (CCPM) stands out more clearly in RMSE, going from 0.397 to 0.366 with manual features and from 0.425 to 0.384 with automatic features. Although using manual features provided the best overall results, the best automatic features model only presents a relative 0.5% lower correlation and 5% higher RMSE than the best manual model.

No features were removed during bidirectional stepwise regression. For LASSO, although it shows good performance, it never provided the best results. Applying PCA to the entire feature matrix provided worse results, although it shows better performance than stepwise selection over raw features in the automatic methods.

Analyzing the features commonly chosen by stepwise regression in the manual case, the 40% selection provided slightly better results, with the selected features being: CCPM (s6) and FastR (s21 – rate of deletions+insertions) from sentences; p1 (CWPM), p2 (WPM) and p25 (Diff1 – difficulty based on pronunciation rules only) from pseudowords. This shows that reading speed of both sentences and pseudowords was relevant, as well as the difficulty of pseudowords based on dubious and infrequent pronunciation rules. The combined rate of deletions and insertions was also chosen, with a negative weight, meaning that although these disfluencies are more common in higher grades, they are often given by fast speakers and this term might appear as a regulatory term to lower their scores that would otherwise be high. For the automatically obtained features selected from stepwise regression, reading speed of pseudowords was not chosen very often, although p22 (DisfR – rate of all disfluencies) does appear in the $>40\%$ selection. Nevertheless, the best model was obtained from the features appearing in more than 5% of the folds, which includes: reading speeds of both sentences and pseudowords (s3-6, p1 and p4), p19 (ExtraR – rate of false-starts+repetitions), p22 (DisfR – rate of all disfluencies) and p25 (Diff1 – difficulty based on pronunciation rules only). There are three common features with the manual analysis (s6, p1 and p25) with the rates of disfluencies being the additions that stand out. Since our automatic method does not detect deletions and insertions, feature s21 (ExtraR) could not be selected for the automatic analysis.

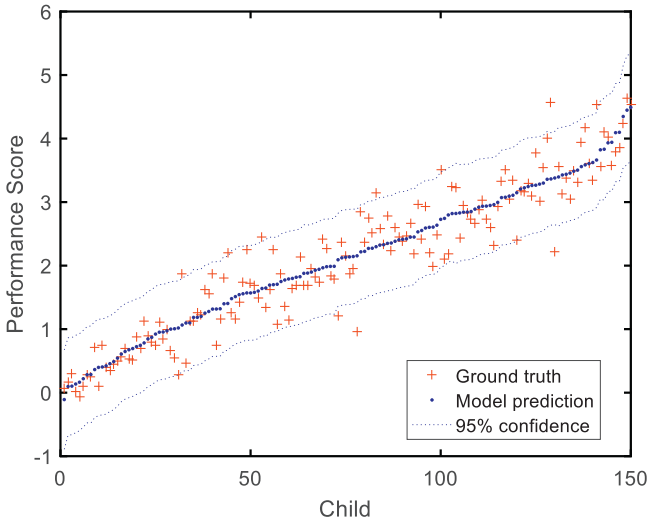

Fig. 11. Per child reference scores (Ground truth) and predicted scores by the best performing Gaussian Process Regression (GPR) model using automatic features, including a 95% confidence interval of the GPR model.

Fig. 11 shows the predicted scores of the best model for automatic features – GPR Sel5% – with their corresponding Ground truth scores, as well as a 95% confidence interval given by the probabilistic GPR model. It can be seen that the GPR model fits most of the reference scores inside its 95% confidence interval, excluding some outliers.

Fig. 12 displays the reference ground truth scores with the predictions of the best model for automatic features (same as above), including the standard deviation (std) of the scores given by teachers for each child (the mean of those values results in the reference score for that child). This deviation, with an average of 0.549, reflects the evaluator uncertainty associated with scoring each child. The RMSE of the predicted scores by the model (0.384) is lower than evaluator std, and most predictions fall inside the deviation interval (again, with some outliers).

Overall, since metrics based on both sentence reading and pseudowords reading tasks were used by the best performing models, it may be concluded that teachers gave their overall impression

Table 7
Performance of multi-feature models on manual and automatic features (with test values after leave-one-out cross-validation).

Manual features			Automatic features		
Model	Corr	RMSE	Model	Corr	RMSE
LR (s6)	0.940	0.397	LR (s6)	0.931	0.425
LR (s6,p5)	0.943	0.386	LR (s6,p5)	0.933	0.419
GPR (s6,p5)	0.948	0.371	GPR (s6,p5)	0.938	0.403
LR all	0.926	0.442	LR all	0.931	0.426
GPR all	0.947	0.375	GPR all	**0.943**	**0.388**
Stepwise add + LR	0.947	0.373	Stepwise add + LR	0.919	0.458
Stepwise add+remove + LR	0.947	0.373	Stepwise add+remove + LR	0.919	0.458
Stepwise add + GPR	**0.949**	**0.367**	Stepwise add + GPR	0.932	0.422
LASSO	0.944	0.387	LASSO	0.932	0.423
LASSO + LR	0.942	0.392	LASSO + LR	0.932	0.421
LASSO + GPR	0.942	0.392	LASSO + GPR	0.939	0.400
PCA 95% + LR	0.917	0.465	PCA 95% + LR	0.916	0.467
PCA 95% + GPR	0.931	0.423	PCA 95% + GPR	0.927	0.436
PCA all + Stepwise + LR	0.909	0.488	PCA all + Stepwise + LR	0.938	0.404
PCA all + Stepwise + GPR	0.939	0.401	PCA all + Stepwise + GPR	0.936	0.408
LR Sel80% (s6,21;p1,25)	0.947	0.373	LR Sel80% (s3,6)	0.937	0.407
LR Sel40% (s6,21;p1,2,25)	0.947	0.373	LR Sel40% (s3,6,21;p22,25)	0.940	0.398
LR Sel5% (s1,6,18,21;p1,2,5,6,25)	0.946	0.377	LR Sel5% (s3,4,5,6,21;p1,4,19,22,25)	0.937	0.405
GPR Sel80% (s6,21;p1,25)	0.949	0.367	GPR Sel80% (s3,6)	0.940	0.397
GPR Sel40% (s6,21;p1,2,25)	**0.949**	**0.366**	GPR Sel40% (s3,6;p22,25)	0.940	0.398
GPR Sel5% (s1,6,18,21;p1,2,5,6,25)	0.947	0.373	GPR Sel5% (s3,4,5,6;p1,4,19,22,25)	**0.944**	**0.384**

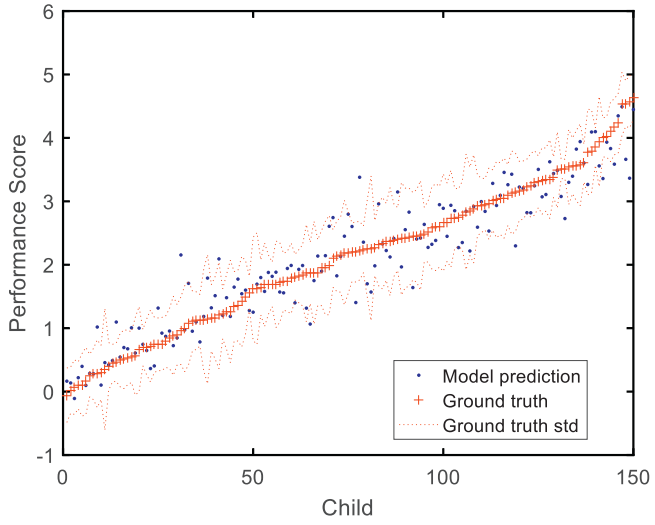

Fig. 12. Per child predicted scores by the best-performing Gaussian process regression (GPR) model using automatic features and reference scores (Ground truth) including the standard deviation (std) interval of the opinion of teachers for each child.

based on both tasks. Although reading speed features were the most important factors for reading ability assessment, detecting disfluencies proves to be of relevance as well, even for rates of specific types of disfluencies, which are different features than when they are considered to compute correct words/syllables/characters per minute. The difficulty of the tasks given is also an important factor to take into account when predicting reading performance and is possibly a normalizing factor. It is reflected in the selection of the difficulty of the pseudowords for both manual and automatic feature models and, indirectly, in the increase in performance obtained when using correct characters per minute instead of correct words per minute, since the length of words may be a measure of difficulty.

5. Conclusion

Aiming to automatically assess the overall reading aloud ability of primary school children, we analyzed reading tasks using sentences and pseudowords as a way to elicit complementary information about their performance. A dataset was carefully designed and collected and several types of reading disfluencies were identified. The average performance of children was shown to increase with grade level on both sentence and pseudoword tasks, although there is a high variation within a grade level. Some of the most common disfluencies were targeted for automatic detection: false starts, repetitions, and mispronunciations. Using task-specific lattices and syllable-based false starts, we managed to detect 80% of disfluency events that result in extra segments, with a false alarm rate lower than 1%. Detecting mispronunciations proved to be much more challenging and, by using a log likelihood measure from the output of a neural network built towards phoneme recognition, we achieved a 5% false alarm rate and 33% miss rate for severe mispronunciations and 45% miss rate when slight mispronunciations are included.

To get ground truth for overall reading aloud score, the opinions of primary school teachers were gathered and their mean opinion taken as reference. Regression models were trained to automatically predict these scores and, although it is undeniable that correct words per minute read in sentences is already a very good measure for what teachers believe the reading level of a child to be, certain features were found to be effective in getting automatic scoring closer to the ground truth. Specifically, features relating to the performance in pseudoword tasks and the difficulty of these tasks were helpful when using either manually or automatically obtained features. Even if all disfluencies were not correctly identified with the automatic methods, the performance of models using features from the automatic annotation to predict overall reading score fell close to the performance based on manual annotation. Gaussian process regression (GPR) models were also shown to perform better than simple linear regression, as they are robust to noise and outliers.

The predicted scores of overall reading performance fell mostly inside the standard deviation of human evaluation, although some outliers are found. Similarly, there are some ground truth scores that are outside the 95% confidence interval of the best GPR model.

Further work needs to investigate which factors that were not considered by the analyzed features lead to these outlier scores.

The developed methods can ideally be applied to a stand-alone application used by teachers and students, where reading tasks are assigned, performed and automatically analyzed, keeping records of a child's performance for multiple tasks over time.

Acknowledgments

This work was supported in part by Fundação para a Ciência e Tecnologia under the project UID/EEA/50008/2013 (pluriannual funding in the scope of the LETSREAD project at Instituto de Telecomunicações). The authors acknowledge the support given by Microsoft to this project. Jorge Proença is with Instituto de Telecomunicações and the Department of Electrical and Computer Engineering and is supported by the SFRH/BD/97204/2013 FCT Grant. We would like to thank the João de Deus, Bissaya Barreto and EBI de Pereira school associations and CASPAE parents association for collaborating in the database collection.

References

Abdou, S.M., Hamid, S.E., Rashwan, M., Samir, A., Abdel-Hamid, O., Shahin, M., Nazih, W., 2006. Computer aided pronunciation learning system using speech recognition techniques. In: Proc. Interspeech 2006. Pittsburgh, USA, pp. 849–852.

Black, M., Tepperman, J., Lee, S., Price, P., Narayanan, S., 2007. Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment. In: Proc. Interspeech 2007. Antwerp, Belgium, pp. 206–209.

Black, M.P., Tepperman, J., Narayanan, S.S., 2011. Automatic prediction of children's reading ability for high-level literacy assessment. Trans. Audio, Speech and Lang. Proc. 19, 1015–1028. doi:10.1109/TASL.2010.2076389.

Bolaños, D., Cole, R.A., Ward, W., Borts, E., Svirsky, E., 2011. FLORA: fluent oral reading assessment of children's speech. ACM Trans. Speech Lang. Process 7 (16), 1–16. doi:10.1145/1998384.1998390, 19.

Buescu, H.C., Morais, J., Rocha, M.R., Magalhães, V.F., 2015. Programa e Metas Curriculares De Português Do Ensino Básico. Ministério da Educação e Ciência.

Candeias, S., Celorico, D., Proença, J., Veiga, A., Perdigão, F., 2013. HESITA(tions) in Portuguese: a database. In: ISCA, Interspeech Satellite Workshop on Disfluency in Spontaneous Speech - DiSS. KTH Royal Institute of Technology, Stockholm, Sweden, pp. 13–16.

Cincarek, T., Gruhn, R., Hacker, C., Nöth, E., Nakamura, S., 2009. Automatic pronunciation scoring of words and sentences independent from the non-native's first language. Computer Speech & Language 23, 65–88. doi:10.1016/j.csl.2008.03.001.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. doi:10.1177/001316446002000104.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, 3rd ed Wiley.

Duchateau, J., Cleuren, L., hamme, H.V., Ghesquière, P., 2007. Automatic assessment of children's reading level. In: Proc. Interspeech 2007. ISCA, Antwerp, Belgium, pp. 1210–1213.

Duchateau, J., Kong, Y.O., Cleuren, L., Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Verhelst, W., hamme, H.V., 2009. Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. Speech Commun. Spoken Lang. Technol. Educ. Spoken Lang. 51, 985–994. doi:10.1016/j.specom.2009.04.010.

Fiscus, J.G., Ajot, J., Garofolo, J.S., Doddington, G., 2007. Results of the 2006 spoken term detection evaluation. In: Proc. SIGIR 2007. Amsterdam, Netherlands, pp. 51–57.

Fuchs, L.S., Fuchs, D., Hosp, M.K., Jenkins, J.R., 2001. Oral reading fluency as an indicator of reading competence: a theoretical, empirical, and historical analysis. Sci. Stud. Read. 5, 239–256.

Hämäläinen, A., Rodrigues, S., Júdice, A., Silva, S.M., Calado, A., Pinto, F.M., Dias, M.S., 2013. The CNG corpus of European Portuguese children's speech. In: Habernal, I., Matoušek, V. (Eds.), Text, Speech, and Dialogue, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 544–551.

Hasbrouck, J., Tindal, G.A., 2006. Oral reading fluency norms: A valuable assessment tool for reading teachers. Read. Teach. 59, 636–644.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer Science & Business Media.

Keuleers, E., Brysbaert, M., 2010. Wuggy: a multilingual pseudoword generator. Behav. Res. Methods 42, 627–633. doi:10.3758/BRM.42.3.627.

Li, X., Ju, Y.-C., Deng, L., Acero, A., 2007. Efficient and robust language modeling in an automatic children's reading tutor system. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 193–196. doi:10.1109/ICASSP.2007.367196.

Liu, Y., Shriberg, E., Stolcke, A., Harper, M.P., 2005. Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. In: Proc. Interspeech. Citeseer, pp. 3313–3316.

Lopes, C., 2012. ChildCAST. Available at http://lsi.co.it.pt/spl/childCAST/ChildCAST_v3.zip.

Lopes, C., Veiga, A., Perdigão, F., 2012. A European Portuguese children speech database for computer aided speech therapy. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Computational Processing of the Portuguese Language, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 368–374.

Medeiros, H., Moniz, H., Batista, F., Trancoso, I., Nunes, L., others, 2013. Disfluency detection based on prosodic features for university lectures. In: Proc. Interspeech. Lyon, France, pp. 2629–2633.

Mendonça, G., Candeias, S., Perdigao, F., Shulby, C., Toniazzo, R., Klautau, A., Aluisio, S., 2014. A method for the extraction of phonetically-rich triphone sentences. In: Proc. of the International Telecommunications Symposium (ITS). São Paulo, Brazil, pp. 1–5. doi:10.1109/ITS.2014.6947957.

Moniz, H., Batista, F., Mata, A.I., Trancoso, I., 2014. Speaking style effects in the production of disfluencies. Speech Commun. 65, 20–35. doi:10.1016/j.specom.2014.05.004.

Mostow, J., Roth, S.F., Hauptmann, A.G., Kane, M., 1994. A prototype reading coach that listens. In: Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1), AAAI '94. American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 785–792.

National Reading Panel, 2000. Teaching Children to read: An evidence-Based Assessment of the Scientific Research Literature On Reading and Its Implications For Reading Instruction. National Institute of Child Health and Human Development, USA.

Pellegrini, T., Hämäläinen, A., de Mareüil, P.B., Tjalve, M., Trancoso, I., Candeias, S., Dias, M.S., Braga, D., 2013. A corpus-based study of elderly and young speakers of European Portuguese: acoustic correlates and their impact on speech recognition performance. In: Proc. Interspeech 2013. Lyon, France, pp. 852–856.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi Speech Recognition Toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, Hilton Waikoloa Village, Big Island, Hawaii, US.

Proença, J., Celorico, D., Candeias, S., Lopes, C., Perdigão, F., 2015. Children's reading aloud performance: a database and automatic detection of disfluencies. In: ISCA - Conf. of the International Speech Communication Association - INTERSPEECH. Dresden, Germany, pp. 1655–1659.

Rasmussen, C.E, Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, Massachusetts.

Rocha, P., Santos, D., 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Proc. PROPOR 2000. Atibaia, São Paulo, Brazil, pp. 131–140.

Santos, A.L. Corpus Santos - European Portuguese ISLRN 532-620-702-768-3. [WWW Document]. URL (accessed 3.10.16).

Santos, A.L., Généreux, M., Cardoso, A., Agostinho, C., Abalada, S., 2014. A corpus of European Portuguese child and child-directed speech. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland, pp. 1488–1491.

Soares, A.P., Medeiros, J.C., Simões, A., Machado, J., Costa, A., Iriarte, Á., de Almeida, J.J., Pinheiro, A.P., Comesaña, M., 2014. ESCOLEX: a grade-level lexical database from European Portuguese elementary to middle school textbooks. Behav. Res. Methods 46, 240–253. doi:10.3758/s13428-013-0350-1.

Speecon Consortium, 2005. Portuguese Speecon Database. ELRA-S0180, ISLRN 824-839-200-501-4.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B (Methodological) 58, 267–288.

Veiga, A., Celorico, D., Proença, J., Candeias, S., Perdigão, F., et al., 2012. Prosodic and Phonetic Features for Speaking Styles Classification and Detection. In: Toledano, D.T., Giménez, A.O., Teixeira, A., Rodríguez, J.G., Gómez, L.H., Hernández, R.S.S., et al. (Eds.), Advances in Speech and Language Technologies For Iberian Languages. Springer, Berlin Heidelberg, pp. 89–98.

Veiga, A., Lopes, C., Sá, L., Perdigão, F., et al., 2014. Acoustic Similarity Scores for Keyword Spotting. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Nunes, M., et al. (Eds.), Computational Processing of the Portuguese Language, Lecture Notes in Computer Science. Springer International Publishing, pp. 48–58.

Yilmaz, E., Pelemans, J., hamme, H.V., 2014. Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model. In: Proc. Interspeech 2014. Singapore, pp. 969–972.