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/© 2017 Elsevier B.V. All rights reserved.

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, sound-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