Automatic assessment of expressive oral reading

Daniel Bolaños a,*, Ronald A. Cole a, Wayne H. Ward a, b, Gerald A. Tindal c, Paula J. Schwanenflugel d, Melanie R. Kuhn e

a Boulder Language Technologies, 2960 Center Green Court, Suite 200, Boulder, CO 80301, USA
b University of Colorado at Boulder, Boulder, CO 80309, USA
c University of Oregon, Eugene, OR 97403-5267, USA
d The University of Georgia, Athens, GA 30602, USA
e Boston University, Boston, MA 02215, USA

Received 8 November 2011; received in revised form 3 August 2012; accepted 7 August 2012
Available online 17 August 2012

Abstract

We investigated the automatic assessment of expressive children’s oral reading of grade level text passages using a standardized rubric. After a careful review of the reading literature and a close examination of the rubric, we designed a novel set of prosodic and lexical features to characterize fluent expressive reading.

A number of complementary sources of information were used to design the features, each of them motivated by research on different components of reading fluency. Features are connected to the child’s reading rate, to the presence and number of pauses, filled-pauses and word-repetitions, the correlation between punctuation marks and pauses, the length of word groupings, syllable stress and duration and the location of pitch peaks and contours.

The proposed features were evaluated on a corpus of 783 one-minute reading sessions from 313 students reading grade-leveled passages without assistance (cold unassisted reading). Experimental results show that the proposed lexical and prosodic features provide complementary information and are able to capture the characteristics of expressive reading. The results showed that on both the 2-point and the 4-point expressiveness scales, computer-generated ratings of expressiveness agreed with human raters better than the human raters agreed with each other. The results of the study suggest that automatic assessment of expressive oral reading can be combined with automatic measures of word accuracy and reading rate to produce an accurate multidimensional estimate of children’s oral reading ability.

© 2012 Elsevier B.V. All rights reserved.

Keywords: Oral reading fluency; Expressive reading; Prosody; Children’s read speech; Education

1. Introduction

Oral reading fluency (ORF) is frequently used, along with other measures, to assess an individual’s reading level and proficiency. ORF is defined as the ability to read a grade level text accurately, at a natural speaking rate, and with proper expression (National Reading Panel, 2000). Of the three components of oral reading fluency—accuracy, rate and expressiveness—accuracy and rate have been most studied. In fact, these two components have been combined into a single measure, Words Correct Per Minute (WCPM), which is widely used to assess individuals’ reading ability.

Speed and accuracy are thought to emerge concurrently as readers develop automaticity in reading (Logan, 1988). The gains in speed are largest early in the development of reading. For example, when WCPM are calculated,
according to the Hashbrouck and Tindal (2006) reading fluency norms (which are widely used by teachers to determine children’s fluency when reading classroom materials), children in 3rd grade gain an average 36 WCPM across the school year, but between 6th and 8th grade, they gain only 1 WCPM. Change in WCPM is a widely accepted marker of growing reading fluency and it is possibly the best single indicator of reading skills in general during elementary school. Currently, most curriculum-based assessments of reading fluency rely on WCPM as the basic indicator of overall reading skill.

Progress monitoring systems such as AIMSweb R-CBM Oral Reading (Shinn and Shinn, 2002), DIBELS Oral Reading Fluency (Good and Kaminski, 2002) and EasyCBM (Alonzo et al., 2006), among others, have shown that measures using WCPM have excellent reliability and good predictive validity with various state-level end-of-year criterion-tests of reading, standardized assessments of reading skill, and reading comprehension (Fuchs et al., 2002; Wayman et al., 2007). However, assessments based on WCPM measure only a child’s current reading rate and accuracy, and not the third element in the National Reading Panel definition of reading fluency, proper expression. From a measurement standpoint, this seriously under-represents the construct of reading fluency, the net effect of which is to de-emphasize reading expressiveness, or prosody, in both assessment and instruction (Kuhn et al., 2010; Schwanenflugel and Benjamin, 2012).

Expressive reading is typically measured in two ways. One way is to have teachers or evaluators use rating schemes that have been created directly for this purpose. The NAEP Oral Reading Fluency Scale is one such measure (Pinnell et al., 1995). This 4-point scale distinguishes reading that sounds primarily word-by-word from reading that occurs in larger, meaningful phrase groups (Pinnell et al., 1995, p. 15). In the development of the scale, accuracy in reading was intentionally kept out of the scale to distinguish mere accuracy from true reading fluency (Pinnell et al., 1995; Daane et al., 2005). The NAEP scale has been used in a variety of research studies (Daane et al., 2005; Kuhn, 2005; Pinnell et al., 1995; Valencia et al., 2010). The most recent NAEP study of over 1000 fourth graders (Daane et al., 2005) found 61% fluent, with 10% achieving the highest rating level. Further, Valencia et al. (2010) used the NAEP scale as a measure of reading prosody along with measures of accuracy and rate, and found that the NAEP accounted for significant variance beyond accuracy and rate in reading comprehension in 2nd, 4th, and 6th grade readers. Another scale, the Multidimensional Fluency Scoring Guide (MFSG) (Rasinski et al., 2009); originally, the Multidimensional Fluency Scale; (Zutell and Rasinski, 1991) has three 4-point scales encompassing phrasing and expression, accuracy and smoothness, and pacing. Like the NAEP, the focus of the MFSG is on prosody or expression. However, of the two, the NAEP is the most widely used. We therefore chose the NAEP expressiveness scale for our research.

A second approach to measuring expressiveness during oral reading examines changes in pitch, rhythm, stress, and pausing present in children’s reading using spectrographic examination of voice features (Benjamin and Schwanenflugel, 2010; Cowie et al., 2002; Schwanenflugel et al., 2004; Miller and Schwanenflugel, 2006; Miller and Schwanenflugel, 2008). Research has shown that, as children become more fluent readers, they develop more expressive and appropriate oral reading prosody when measured using spectrographic measures. As children become fluent readers, they begin to make shorter and less variable inter-sentential pauses, shorter and less frequent intra-sentential pauses, larger pitch declinations, and they come to display a more adult-like intonation contour and more linguistically relevant pausing (Benjamin and Schwanenflugel, 2010; Clay and Imlach, 1971; Cowie et al., 2002; Miller and Schwanenflugel, 2006; Miller and Schwanenflugel, 2008). Longitudinal changes in reading prosody between first and second grade have been found to be predictive of later reading fluency, beyond early reading rate measures (Miller and Schwanenflugel, 2008). Children who have good reading prosody also tend to have better reading comprehension than would be expected on the basis of their reading rates alone (Benjamin and Schwanenflugel, 2010; Miller and Schwanenflugel, 2006; Miller and Schwanenflugel, 2008; Schwanenflugel and Benjamin, 2012).

The focus of our research is to demonstrate the feasibility of automating assessment of oral reading fluency, including accurate estimates of accuracy, rate and expressiveness. Below, we discuss the many benefits to teachers and students if this goal can be achieved. In our previous research (Bolanos et al., 2011) we introduced FLORA (FLuent Oral Reading Assessment), a web-based system that uses automatic speech recognition (ASR) to estimate the number of words correct per minute produced by first through fourth grade students who read grade-level text passages out loud for one minute. Relative to human raters, FLORA was shown to be effective at measuring reading rate and accuracy (WCPM).

In the current study, we present research that extends FLORA to fully automated assessment of expressive reading, in addition to WCPM. Automatic assessment of ORF has many potential benefits such as saving teachers’ time (that can be used for reading instruction). At the Boulder Valley School District, the site of our study, elementary school teachers average about 6 schools days each year assessing their students’ reading proficiency. Data from automated assessments, including digitized recordings, can be entered into a database for all student assessments, enabling teachers to review progress reports for individual students as well as to listen to samples read aloud across successive assessments. Data could also be analyzed, summarized, and displayed to answer questions about changes in students’ reading abilities for classrooms and schools within and across school districts. Additionally, the automatic assessment system proposed is deterministic, being
We defined and developed an initial baseline set of lexical and prosodic features that we expected to correlate well with the fluency level of the speaker. These features are described in more detail below. Lexical features include WCPM, n-gram back-offs and number of repetitions. Prosodic features, intended to provide complementary information that can help to discriminate the speaker’s expressiveness, include the location of the pitch accent, word and syllable durations, filled and unfilled pauses and punctuation marks. Features were extracted using an automatic speech recognition (ASR) system used in both recognition and forced-alignment modes, a speech activity detection (SAD) module, a pitch extractor (Sjölander, 1997–2004) and a syllabification toolkit (Fisher, 1996).

The proposed features were evaluated on a corpus of 783 one-minute recordings of first through fourth grade students reading grade level text passages. Each of the recordings was scored from 1 to 4 using the NAEP fluency scale by at least two independent scorers with experience assessing reading proficiency. The scorers listened to each 60 s story in 20 s intervals, and provided a 1–4 rating for each interval. Finally they attached a global NAEP score to the recording based on the NAEP scores assigned to each 20 s segment.

The use of prosodic features in automatic speech recognition has been studied for a variety of tasks including automatic sentence boundary detection (Liu et al., 2006), emphasis detection (Brenier et al., 2005) and improving speech recognition (Vicsi and Szaszaĥ, 2010). In the particular case of oral reading assessment of children’s speech, previous work has already shown the effectiveness of prosodic features. In (Patel and McNab, 2011) prosodic features like word duration, peak intensity and peak F0 were used to evaluate the effectiveness of a tool to enhance expressive oral reading.

The automatic assessment of expressive oral reading has been investigated by Mostow and his collaborators (Mostow and Duong, 2009; Duong and Mostow, 2010; Duong et al., 2011). In (Duong et al., 2011) two alternative methods of measuring the prosody during children’s oral reading were described and compared. The first method consisted of generating a prosodic template model for each sentence in the text. The template was based on word-level features like pitch, intensity, latency and duration extracted from fluent adult narrations. Children’s expressive reading was then scored based on the similarity between the prosodic realization of each sentence read aloud and the template constructed from an adult reading the same sentences.

The second method investigated adult narrations to train a general duration model that could be used to generate expected prosodic contours of sentences for any text, so an adult reader was no longer required to generate sentence templates for each new text. Both methods were evaluated by their ability to predict student’s scores on fluency and comprehension tests, and each produced promising results.
with the second, automated method for generating prosodic sentence templates outperforming adult narrations of each individual text. However none of these methods could satisfactorily classify sentences according to a standard fluency rubric, which was probably due to the low human inter-rater reliability.

The work presented here differs from Duong et al. (2011) in several of ways. Our assessment of children’s prosodic reading does not make use of adult speech as a reference, instead we propose a machine learning approach to learn directly from an annotated corpus of children read speech. In addition to prosodic features like duration or pitch we also used lexical features and features based on filled pauses and syllabic information.

Moreover, in (Duong et al., 2011) the evaluation data was presented to the student one sentence at a time and feedback was provided to the student when the system detected reading difficulties or when the student requested it. In our study, we evaluated expressiveness of one minute recordings of text passages read aloud by students who did not receive assistance or feedback while reading; thus the expressiveness scores in our study were generated through feature extraction and classification of speech across an entire text passage, rather than sentence by sentence. In (Duong et al., 2011) the Multidimensional Fluency Scale (MFS) (Zutell and Rasinski, 1991) was used to assess sentence-level fluency. Results showed that inter-rater reliability on this scale—which comprises four 4-point scales encompassing expression, phrasing, smoothness and pace—was unsatisfactory, especially when used at the sentence level. In our work, instead of using the MFS, we used the NAEP Oral Reading Fluency Scale (Pinnell et al., 1995) which has a single dimension and is much simpler to administer. In addition to that, we administered it for each one-minute speech recordings instead of for each sentence. Our evaluation was also carried out on many more speakers than the evaluation in (Duong et al., 2011) so our results may generalize better.

The remainder of the article is organized as follows: In Section 2 we introduce the prosodic features used in this study as well as other features that correlate with oral reading fluency. Section 3 deals with the data collection and scoring. Section 4 describes the experimental setup and the experiments carried out. Section 5 presents the results. Finally, Section 6 presents the discussion, conclusions and future work.

2. Proposed features

In this section we introduce the proposed features for automatic assessment of expressive reading. All the proposed features were measured and used to assign a classification score to each one-minute recorded story. The features are text-independent, and can be applied to any recorded text to assign an expressive score from 1 to 4 on the NAEP scale.

2.1. Lexical features

The features listed below were obtained from the 1-best hypothesis generated by the speech recognition system.

- **L1**: Words Correct Per Minute. This feature is the number of words correctly read by the speaker during the reading session, it expresses the student’s reading rate. Words are scored as incorrect if they are skipped over or scored as misread by the recognizer. 

- **L2**: Total number of words spoken by the student during the one-minute reading session. This value is normalized across recordings since all of them are one minute in length. This feature, which measures the child’s speech rate, i.e., all of the words identified as spoken by the recognizer while reading the text, differs from WCPM, which measures the number of words read correctly, but does not count insertions or misread words.

- **L3**: Number of word repetitions: we count the word repetitions found in the hypothesized word string produced by the recognizer that are not in the reference text (the passage read). In this work we address only one and two-word repetitions, which accounted for more than 90% of the total number of repetitions. According to the NAEP scale the number of repetitions correlates negatively with the reader’s fluency level. 

- **L4**: Number of trigram back-offs: Word co-occurrences are represented during word recognition as sequences of trigrams. Trigrams are computed automatically for each text passage based on the words in the passage. The number of times a trigram in the speech recognition hypothesis is not found can thus provide a measure of fluency. A trigram back-off occurs, for example, when the speaker decodes a word incorrectly (which causes the speech recognition system to hypothesize a different word) or inserts a word.

- **L5**: Variance of the sentence reading rate: a fluent reader should read all sentences in the passage at a similar reading rate regardless of their difficulty. A less fluent reader will have more difficulty with individual words in some sentences, which will produce higher variance across
sentences. Benjamin and Schwanenflugel (2010) found that intrasentential pausing correlates with word decoding difficulties.

2.2. Prosodic features

Prosodic features measure speech behaviors related to the discourse structure of a text, such as phrasing, emphasis, new information, and tone. These behaviors are realized as changes in pitch across words and changes in the duration and amplitude of syllables within words, as well as the expression of emotional states consistent with the discourse and syntactic structure of the text. A total of 15 prosodic features were proposed, which design was based on a review of the literature on acoustic-phonetic and prosodic analysis of read speech (Kuhn et al., 2010).

Features P1, P2, P3 and P4 aim at expressing whether the child is paying attention to punctuation. Research indicates that understanding the role and purpose of punctuation marks is an important component of fluent reading (Rasinski, 2004). We decided to distinguish between pausing and pitch-based behaviors at sentence-final and within-sentence punctuation marks since they may discriminate between more and less fluent readers. According to Miller and Schwanenflugel (2006) fluent readers do not necessarily pause at all commas. Interestingly, research indicates that within-sentence pauses at punctuation marks like commas are rare among adults Chafe, 1988 and fluent child readers (Miller and Schwanenflugel, 2006). The presence of such events typically occurs for very long sentences where the reader might need to take a breath, (which is required more often for slow readers) (Benjamin and Schwanenflugel, 2010).

Features P5, P6, P7 and P8 are related to the number and duration of pauses made during reading. They are also connected to the length of word-groupings which plays a major role in the definition of the NAEP levels. According to Clay and Imlach (1971), the best readers use few pauses while the poorest readers pause longer and more often. A significant number of pauses can be a symptom of word-by-word reading.

Features P9, P10 and P11 are related to the number and duration of filled pauses. The number of filled pauses is correlated to the decoding abilities and overall fluency. According to the NAEP scale, hesitations are rare in fluent readers.

Features P12 and P13 aim to express whether the reader is over extending the duration of syllables. Lengthened syllables like ‘hisss’ or ‘theeee’ play a role similar to filled pauses and thus can be an indicator of poor fluency.

Features P14 and P15 aim to capture correlations between stress or emphasis given to certain syllables within a sentence, which is an important component of prosody. In continuous speech, only some of the stressed syllables within words (lexical stress) are actually stressed; that is, a speaker will typically modify the prosodic contour of a sentence to stress words in order to convey new and/or important information. Words that are stressed in this way are typically longer in duration and are characterized by a pitch accent or pitch peak. Thus, pitch accent is a more reliable estimator of word and syllable stress than lexical stress.

In the English language, pitch accent is characterized by pitch movement, vowel duration and increased energy. We are aware that the correlation between predicted pitch accent and higher pitch or longer duration should be higher than the correlation between these features and lexical stress. The goal in designing this feature was to discriminate more fluent readers from less fluent ones, assuming that more fluent readers will have better comprehension of the text and will place the pitch accent always on a lexically stressed syllable during oral reading. However, predicting pitch accent is a complex task by itself and given the limited amount of data available for training and evaluation purposes and the availability of lexical stress markers in pronunciation dictionaries, we considered these features reasonable. Basically, we are saying that fluent readers will have more pitch-accented words than less fluent readers, since they are constructing meaning as they read which may be reflected in their prosody as pitch peaks within words.

Note that feature P14 is the only pitch-based feature proposed and that it is text-independent. Text-dependent pitch-based features have been proposed in the literature to assess prosody (Mostow and Duong, 2009), however, our intent is to build a feature extraction method that works on any text.

- **P1**: Ratio between the number of times the speech activity detection module hypothesizes a silence region (or breath noise) at a sentence-ending punctuation mark (period, question mark or exclamation point) and the total number of sentence-ending punctuation marks in the text. According to the NAEP scale a fluent reader should read in meaningful phrase groups. The phrasing should preserve the author’s syntax.
- **P2**: Ratio between the number of times a silence region (or breath noise) is hypothesized at a within-sentence punctuation mark (comma, colon and semicolon) and the total number of within-sentence punctuation marks in the text. This feature is analogous to P1, but may reflect a more sophisticated processing of the text by the reader.
- **P3**: Average length of silence regions aligned to punctuation marks.
- **P4**: Ratio between the length of silence regions (and breath noise) aligned to punctuation marks and the total length of non-speech (silence, breath and filled pauses) regions in the recording. Word-by-word reading is a characteristic of poor readers, if a reader stops at every word boundary they will be stopping at every punctuation mark too. This feature attempts to separate syntac-
tically motivated pauses from pauses caused by time spent decoding individual words. A word-by-word reader should be assigned to the first level in the NAEP scale.

- **P5**: Average number of words between two silence regions (or breath noise). This feature is expected to correlate with the length of word-groupings, which is a good indicator of reading fluency. According to the NAEP scale, students assigned to the third level “read primarily in three- or four-word phrase groups” while students assigned to the first level read “primarily word-by-word”.

- **P6**: Number of silence regions. Total number of times the reader pauses.

- **P7**: Average duration of silence regions.

- **P8**: Duration of the longest silence region. Long silence regions may indicate problems decoding some words.

- **P9**: Number of filled pauses.

- **P10**: Average duration of filled pauses.

- **P11**: Duration of the longest filled pause.

- **P12**: Average syllable length.

- **P13**: Maximum syllable length.

- **P14**: Ratio between the average pitch (F0) of lexically stressed syllables and unstressed syllables.

- **P15**: Ratio between the average duration of lexically stressed syllables and unstressed syllables.

### 2.3. Feature extraction

The discrimination between silence and speech regions used a Speech Activity Detection (SAD) system. The SAD system consisted of two 5-state Hidden Markov Models (left-to-right without skip), one to model silence and another to model speech. About 20 mixture components for both speech and silence were tied across the respective five states. A penalty parameter is used to control the detection of silence regions within the audio.

Punctuation-related features were measured by performing an automatic alignment between the hypothesis from the speech recognition system and the text passage, and then identifying the location of the punctuation marks in the text passage into the recorded speech. Then, the SAD system was used to detect whether punctuation marks and silence regions (or the symbol expressing breath-noise) overlap in the time aligned hypothesis. The hypothesized location and number of filled pauses within the recordings was obtained from the 1-best hypothesis from the recognizer, which contains word and word-level alignments. Filled pauses as well as breath-noise and other non-speech events were modeled explicitly in the automatic speech recognition (ASR) system through the use of specialized phonetic symbols, as described in Section 4.1. Phone and syllable boundaries needed to compute features connected to syllable duration are obtained from aligning the recognition hypothesis to the audio. The syllabification of the lexicon was done using the NIST syllabification software (Fisher, 1996).

Pitch and formant extraction were carried out using the Snack Speech Toolkit Sjölander, 1997–2004.

Finally, we note that all these features were extracted on same length recordings so no duration normalization was necessary.

### 3. Data collection and scoring

#### 3.1. Data collection

Features described in Section 2 were evaluated on 783 recordings of text passages read aloud by 313 first through fourth grade students in four elementary schools in the Boulder Valley School District (BVSD) in Colorado. School 1 had 53.8% students receiving free or reduced lunches, and the lowest literacy achievement scores of the three schools on Colorado state literacy test given to third grade students; 53% third grade students in School 1 scored proficient or above on the state reading assessment. School 2 had 51.7% students with free or reduced lunch (similar to School 1), but 79% of third grade students tested as proficient or above on the state literacy test. School 2 was a bilingual school with nearly 100% English learners (ELs) who spoke Spanish as their first language. School 3 had 18.4% of students with free or reduced lunch, 85% of students were proficient or above in the state literacy test. School 3 also had relatively few ELs.

The 783 recordings yielded approximately 13 hours of speech data. Data were collected from students in their classrooms at their schools. Our project staff took up to three laptops to each school, and recorded speech data from all students in each classroom. The data was collected using the FLORA system (Bolanos et al., 2011), which was configured to enroll each student, and then randomly select one passage from a set of 20 standardized passages of similar difficulty at the student’s grade level. Depending upon the number of students that needed to be tested on a given day, each student was presented with two or three text passages to read aloud.

During the testing procedure, the student was seated before the laptop, and wore a set of Sennheiser headphones with an attached noise-cancelling microphone. The experimenter observed or helped the student enroll in the session which involved entering the student’s gender, age and grade level. FLORA then presented a text passage, started the one minute recording at the instant the passage was displayed, recorded the student’s speech and relayed the speech to the server.

Because testing was conducted in May, near the end of the school year, classroom teachers had recently assessed their student’s oral reading performance (using text passages different from those used in our study). About 20% of the time, teachers requested that specific students be presented with text passages either one or two levels below or one or two levels above the student’s grade level. Thus, about 80% of students in each grade read passages at their grade level, while 20% of students read passages above or
below their grade level, based on their teachers’ recommendations. Table 2 summarizes the corpus used to evaluate the features proposed in this study.

Twenty text passages were available for reading at each grade level. The standardized text passages were downloaded from a website (Good et al., 2007) and are freely available for non-commercial use. The twenty passages were designed to be about the same level of difficulty at each grade level, and were designed specifically to assess oral reading fluency. Oral reading fluency norms have been collected for these text passages for tens of thousands of students at each grade level in fall, winter and spring semesters, so that students can be assigned to percentiles based on national WCMP scores (Hasbrouck and Tindal, 2006). Table 3 provides statistics of stories at each grade level.

### 3.2. Human scoring of recorded stories

In order to evaluate the ability of the proposed features to classify recordings according to the NAEP scale, each of the one-minute recordings collected was scored independently by at least two former elementary school teachers. A set of 70 stories of the total 783 stories were scored by the five available teachers while the other recordings were scored by just two of them, which were randomly assigned to each scorer. Table 4 shows the number of recordings scored by each teacher. A training session was held before the teachers independently scored the recordings. At the beginning of the training session, the NAEP scoring instructions were reviewed. The teachers then listened to individual stories as a group, in 20 second intervals, and rated each interval on the 1–4 NAEP scale. Each teacher was then shown how to combine the three scores to arrive at a single score for each story. Teachers were asked to use their best judgment to combine the three scores according to their perception of overall fluency across the whole session rather than using deterministic methods like taking the mean or mode of the three scores. The teachers were then able to review the story, and discuss their scores.

The scorers used a tool that:

1. Retrieved a recorded story (i.e. a one-minute reading session) from the corpus. Stories were retrieved randomly among those yet to score. Stories were removed from selection as soon as the desired number of teachers scored each recording.

2. Enabled the scorer to listen to the three consecutive 20 second segments of the recording and replay any portion of each segment while viewing the story text. The segments were presented in order, with each segment assigned a score of 1, 2, 3 or 4 based on the NAEP scale before the next segment was presented.

3. Enabled the scorer to attach a global NAEP score to the whole recording. During this stage the NAEP scores for each of the three segments were shown to the scorer.

4. Enabled the scorer to label the recorded speaker as native or non-native. Which resulted in about 89% of the recorded speakers labeled as native.

5. Save the scores into a database.

Thus, the following information was generated for each recording:

- The NAEP score for each of the three consecutive 20 seconds of audio.
- The overall NAEP score.
- Whether the speaker was a native or a non-native.

The procedure of dividing the recording into three consecutive 20 s segments which were scored independently using the NAEP scale was suggested by Drs. Paula Schwanenflugel and Melanie Kuhn (Schwanenflugel and Kuhn, 2011), who have researched expressive reading for over two decades. The idea behind this procedure was to prevent scorers from assigning NAEP scores based on just initial impressions formed at the beginning of the recording or recency effects based on the last part of the recording. By looking at the scores assigned to each segment before assigning the final score, the scorer was expected to develop a better estimate of how fluently and expressively the child read each passage. In addition, the judges agreed that scoring the passages in 20 s segments was convenient and effective. The two reading experts (Dr. Schwanenflugel and Dr. Kuhn) also provided us with independent NAEP

---

**Table 2**

Summary of the data used for the evaluation.

<table>
<thead>
<tr>
<th>Grade</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Recordings</td>
<td>132</td>
<td>259</td>
<td>165</td>
<td>227</td>
<td>783</td>
</tr>
<tr>
<td>#Schools</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>#Students</td>
<td>53</td>
<td>104</td>
<td>66</td>
<td>90</td>
<td>313</td>
</tr>
<tr>
<td>#Hours of audio</td>
<td>2:12′</td>
<td>4:19′</td>
<td>2:45′</td>
<td>3:47′</td>
<td>13:03′</td>
</tr>
</tbody>
</table>

**Table 3**

Statistics of the stories for each grade.

<table>
<thead>
<tr>
<th>Grade</th>
<th>#Words</th>
<th># Unique words</th>
<th># Sentences</th>
<th># Words per sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>223</td>
<td>110</td>
<td>25</td>
<td>8.92</td>
</tr>
<tr>
<td>2nd</td>
<td>249</td>
<td>121</td>
<td>21</td>
<td>11.86</td>
</tr>
<tr>
<td>3rd</td>
<td>255</td>
<td>127</td>
<td>20</td>
<td>12.75</td>
</tr>
<tr>
<td>4th</td>
<td>381</td>
<td>164</td>
<td>28</td>
<td>13.61</td>
</tr>
</tbody>
</table>

**Table 4**

Number of recordings scored by each teacher.

<table>
<thead>
<tr>
<th>Scorer</th>
<th># Recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₁</td>
<td>361</td>
</tr>
<tr>
<td>H₂</td>
<td>181</td>
</tr>
<tr>
<td>H₃</td>
<td>488</td>
</tr>
<tr>
<td>H₄</td>
<td>589</td>
</tr>
<tr>
<td>H₅</td>
<td>157</td>
</tr>
</tbody>
</table>
4. Experiments

We conducted a series of experiments to evaluate the ability of the proposed features to correlate with how expressively children read grade level passages, relative to human scorers. In particular, we trained a set of classifiers using the proposed features in order to assign recordings to the different levels on the NAEP scale. The corpus used for the evaluation was described in Section 3.

4.1. Speech recognition setup

Automatic speech recognition was carried out to generate word-level hypotheses and phone-level alignments from the recordings, which were used for feature extraction. Three different speech corpora were used to train the acoustic models of the recognizer used in this work. The University of Colorado Read and Summarized Stories Corpus (Cole and Pellom, 2006) (325 speakers from 1st to 5th grade), the CU Read and Prompted Children’s Corpus (Cole et al., 2006) (663 speakers from Kindergarten through 5th grade) and the OGI Kids’ Speech Corpus (Shobaki et al., 2000) (509 speakers from 1st to 5th grade). A total of 106 hours of speech from these corpora was used to train the acoustic models. Only read speech from the corpora was used.

The speech recognizer used for the development of FLORA is a large vocabulary continuous speech recognition (LVCSR) system developed by Bolanos (submitted for publications). Acoustic modeling is based on Continuous Density Hidden Markov Models (CD-HMMs). Models used for this work were estimated under the Maximum Likelihood (ML) criterion. The accuracy of this system has been evaluated on different tasks comprising read and conversational speech tasks and found to be comparable to that of other state-of-the-art ASR systems (Bolanos, submitted for publications).

The speech data were parameterized using Mel-Frequency Cepstral Coefficients (MFCC). For every speech utterance, 39-dimensional feature vectors, consisting of 12 MFCCs and energy plus first and second order derivatives, were extracted. Additionally, cepstral mean subtraction was applied for noise robustness. Acoustic models were trained under the Maximum Likelihood criterion using Baum-Welch re-estimation. During the Gaussian-splitting stage, two Gaussian components were added to each mixture after each re-estimation iteration. The accuracy of models resulting from each iteration was monitored using a development set. The final acoustic models were composed of a total of 8072 tied triphones and 143k Gaussian distributions. The phonetic symbol set consisted of 50 symbols, plus silence and seven filler-symbols that were utilized to match filled pauses and non-speech events. The optimal insertion penalty for filler and standard symbols was computed over a set of lattices.

A trigram language model was trained for each of the stories using the CMU language modeling toolkit (Rosenfeld, 1994). The speech recognizer used a static decoding network organized as a prefix-tree and then was minimized using a forward-backward merging of nodes. A real time factor of about 0.3 was achieved under very wide decoding beams, due to the small vocabulary size and the use of a trigram level language model look-ahead. The processing time estimate refers to the generation of the 1-best hypothesis, and does not include the feature extraction.

Thanks to the trigram level language model look-ahead and the reduced vocabulary size, a real time factor of about 0.3 was achieved under very wide beams. This includes only the generation of the 1-best hypothesis, not the feature extraction.

Unsupervised adaptation of the means and variances of each of the Gaussian distributions was carried out using Maximum Likelihood Linear Regression (MLLR) before the second recognition pass. A regression tree was used to cluster the Gaussian distributions, it comprised 50 base-classes and the minimum occupation count to compute a transform was set to 1000 feature frames (10 s of speech). Expectation-Maximization clustering was used to cluster the Gaussian distributions according to their means. In addition to MLLR adaptation we applied Vocal Tract Length Normalization (VTLN) in order to compensate for the difference in the length of the vocal tract across speakers. VTLN was applied both during training and decoding.

4.2. Classification

In order to classify the recordings using the features proposed in Section 2, we used the LibSVM implementation of Support Vector Machine classifiers (SVMs) (Chang and Lin, 2001). SVM classifiers (Vapnik, 1995) are a well-established machine learning approach that, due to their remarkable generalization performance, have attracted much attention and gained extensive application in many fields including speech processing. Unlike other traditional techniques like Artificial Neural Networks (ANNs), SVMs perform both an empirical and a structural risk minimization over the training set, resulting in better generalization. One of their limitations is that, in their original formulation, they are binary classifiers so multiple SVM classifiers are typically used to deal with multiclass classification tasks. In our study different classification schemes were adopted to label each sample according to the NAEP scale.
According to the NAEP scale, speakers can be assigned to one of four levels (1, 2, 3 or 4) or can also be classified as non fluent readers (level 1 or level 2) or fluent readers (level 3 or level 4). From the point of view of automatic classification, assigning recordings to different levels in the NAEP scale can be seen as a broad classification task \{1,2\} vs \{3,4\} and two sub-classification tasks 1 vs 2 and 3 vs 4. For the two class separation task (fluent vs non fluent), which corresponds to discriminate between levels \{1,2\} and \{3,4\} in the NAEP scale, we trained a binary SVM. For the four classes separation task, where each class corresponds to a different fluency level according to the NAEP scale, we adopted two alternative classification strategies: one-vs-rest and a Decision Directed Acyclic Graph (DAG) (Platt et al., 2000).

The one-vs-rest strategy is a well-known strategy for dealing with multiclass classification problems in which \(N\) binary classifiers are trained \((N = \#\text{classes})\) to separate the samples of each class from the rest. During the evaluation the output of the \(N\) classifiers needs to be combined in order to label each sample. Thus, we trained a 1-vs-rest classifier for each of the four NAEP levels, each of them using all the samples available in the training partition. The decision making step was based on attaching probabilistic values (Platt, 1999) to the predicted labels and keeping the class with the highest probability. The DAG approach makes sense conceptually because that is essentially what the NAEP scale does, it distinguishes fluent from disfluent speech and then makes finer distinctions.

To implement the DAG strategy we trained three classifiers. The first classifier was trained on samples from all classes and separated samples from classes \{1,2\} and \{3,4\}. This classifier was placed at the root of the tree while two classifiers trained on samples from classes \{1,2\} and \{3,4\} respectively were placed on the leaves of the tree to make the finer grain decisions. To evaluate a sample the tree is traversed from the root to a leaf and the sample is labeled with the class attached to the reached leaf. The advantage of the DAGs method respect to the 1-vs-rest is that it produces more balanced training sets (# positive samples \(\approx\#\) negative samples) which is a desirable property, however it does not take advantage of all the data to train each classifier. In addition, errors made on the first decision layer are unrecoverable.

In all cases, we used 5-fold cross-validation accumulating the results afterwards. Specifically, we divided the corpus into 5 balanced speaker-disjoint subsets. One different subset is kept for testing in each fold, while the remaining subsets are used to train the classifiers. A linear kernel was used to train all the SVMs, the optimal value of the \(C\) hyperparameter (cost attached to misclassifying a training sample) was estimated for each fold independently, using cross-validation on the training partition. Before training and evaluating the classifiers, all the feature values were scaled to the interval \([-1,1]\).

Given that all the samples were annotated by at least two human annotators, there were multiple labels available for each sample. In order to take advantage of this additional information we tried a multi-label training approach in which samples that receive different labels appear multiple times in the training set. Each training sample is weighted in order to express the degree of consistency of the annotators. If two annotators label a sample with the same label, the sample receives a weight of 1.0 and appears just once in the training set, while if the sample is labeled differently it appears twice in the training set with different labels and a 0.5 weight.

### 4.3. Feature analysis

When multiple features of different nature are used within a classification scheme, it is useful to determine the degree to which those features contribute to the classification accuracy of the classifier. When linear SVMs are used, like in the proposed work, it is possible to determine the relevance of the features by looking at the weight vector \(w \in \mathbb{R}^n\) in the decision function of the SVM \(D(x) = w \cdot x + b\) (Guyon et al., 2002). The larger the absolute value \(|w_j|\) of a feature \(j\) is, the more important role the feature plays in the decision function. This method is straightforward and has shown good results in the literature (Guyon et al., 2002; Chang and Lin., 2008). Once the linear SVM is trained we extracted the vector of weights \(w\) using tools found in (Chang and Lin, 2001).

### 5. Results

In this section we analyze the features proposed in Section 2 and the results of the classification approach described in Section 4.2 for assessing expressive oral reading. First we analyze the classification accuracy for the lexical and prosodic features proposed. We then analyze the agreement and correlation between human scores and the proposed automatic scoring system for the NAEP scale.

#### 5.1. Recognition results

Recognition results from the ASR system built were measured in terms of the word error rate metric (WER). The WER was computed by counting the number of edit errors between the transcription and the hypothesis and dividing by the number of words in the transcription. Given that most lexical and prosodic features were extracted using the one-best recognition output, minimizing the WER is expected to have a direct impact on reducing the noise in the feature extraction and thus improve the classification accuracy. The WER after VTLN and MLLR speaker adaptation was 10.7%.

#### 5.2. Classification accuracy

Table 5 shows the classification accuracy of the proposed features. Classification accuracy was computed on the corpus of 783 recordings described on Section 3. Each
of the recordings was labeled using the classification approaches described in Section 4.2 and the score produced by the classifiers is compared to scores from all the available human raters, as described in Section 3.2. We note that there exists an upper bound to the classification accuracy that can be attained by the classifier. The reason is that whenever the human raters score the same recording differently, which occurred 31% of the time, there is an unrecoverable classification error. (Recall that we implemented multi-label training but single-label classification.) Classification results using the 1-vs-rest strategy were significantly worse compared to those obtained using the DAGs approach, thus we only report results from the later. We attribute this result to the fact that the 1-vs-rest strategy produces highly unbalanced datasets in which often the best classification consists of labeling all the samples with the label of the dominant class. We tried to overcome this problem by weighting classification errors differently on both sides during the training stage, but this strategy did not improve performance.

Examination of Table 5 reveals that both lexical and prosodic features contribute similarly to the classification accuracy for the NAEP-2 task. This can be initially considered an unexpected result since lexical aspects like the number of words read correctly are expected to play a major role when discriminating between fluent and non-fluent readers. However it is important to note that some of the prosodic features defined in this study are very correlated to the lexical features. For example, it is obvious that the number of words correctly read in a one-minute reading session (lexical feature L1) is correlated to the average number of words between two silence regions (prosodic feature P5), the number of silence regions (prosodic feature P6), the average duration of a silence region (prosodic feature P7) or the number of filled pauses (prosodic feature P9).

For the NAEP-4 tasks, lexical features seem to have a dominant role. We attribute this to the feature L1 (Words Correct per Minute), which by itself provides a 71.78% accuracy for the NAEP-4 task. As expected the automatically computed WCPM, which comprises two of the three reading fluency cornerstones (speed and accuracy) plays a fundamental role. An interesting observation is that errors present in the estimation of the WCPM scores did not have a negative effect on the classification accuracy. We know from Bolanos et al. (2011) that WCPM scores computed by FLORA and used in this work to extract the feature B2 differ by about 3.5 words compared to the human scorers, who differed from each other by slightly less than two words on average across all text passages. In order to understand possible effects of the more variable automatic estimate of WCPM, we conducted a “cheating experiment” in which we substituted WCPM scores produced by human scorers for the automatically extracted L1; this experiment revealed no difference in the results. This indicates that the classifier is robust to small errors in the estimation of the WCPM scores.

For both tasks lexical and prosodic features seem to provide complementary information that leads to an improved classification accuracy when combined. Multi-label training provided small improvements in performance.

In particular, for the NAEP-4 classification task the classification accuracy for the 1 vs 2 and 3 vs 4 classifiers is 76.27% and 75.86% respectively. This classification accuracy is computed considering samples that were misclassified in the root of the tree ([1,2] vs [3,4] classifier) and thus could not be correctly classified by the lower layer classifiers.

Fig. 1 shows the distribution of recordings across the NAEP levels according to humans and the machine. In the case of humans the distribution is normalized by the number of recordings given that there were multiple labels for each recording. It can be seen that both distributions are very similar which is a desirable property.

### 5.3. Inter-rater agreement and correlation

In this section we show inter-rater agreement and correlation results for the best system from the previous section (multi-label training using all the features) compared to the human scorers.

Table 6 shows the inter-rater agreement for the tasks of classifying recordings into the broad NAEP categories (fluent vs non fluent), referred as NAEP-2, and the 4 NAEP categories, referred as NAEP-4. For the NAEP-2 task the
inter-rater agreement is measured using the Cohen’s Kappa coefficient ($\kappa$) (Cohen, 1960) which is more robust than simple percent agreement because it takes into account the chance agreement; where $p(a)$ is the probability of observed agreement while $p(e)$ is the probability of chance agreement.

For the NAEP-4 task we measured the inter-rater agreement using the Weighted Kappa coefficient ($\kappa$) (Cohen, 1968) which is more suitable for ordinal categories given that it weights disagreements differently depending on the distance between the categories (we used linear weightings). As a complementary metric for this task we have computed the Spearman’s rank correlation coefficient (Spearman, 1904). In a number of classification problems, like emotion classification, the data is annotated by a group of human raters who may exhibit consistent disagreements on similar classes or similar attributes. In such classification tasks it is inappropriate to assume that there is only one correct label since different individuals may consistently provide different annotations (Steidl et al., 2005). While the NAEP scale is based on clear descriptions of reading behaviors at each of four levels, children’s reading behaviors can vary across these descriptions while reading, and individuals scoring the stories may differ consistently in how they interpret and weight children’s oral reading behaviors. For this reason, we believe that examining correlations between human raters and between human raters and the machine classifiers is a meaningful and useful metric for this task.

Each row in the table shows the agreement and correlation coefficients of each rater respect to the other raters, which are the other human raters in the case of a human rater (denoted by $H$) or all the human raters in the case of the automatic scoring system (denoted with a $M$ as in Machine). Recall (see Table 4) that not all the raters scored the same number of recordings.

Each row in Table 6 displays the agreement and correlation coefficients of each rater with respect to the other raters. Human raters ($H$) are compared to other human raters, and the automatic scoring system (denoted with an $M$ for machine) is compared to the other human raters. Recall (see Table 3) that not all the raters scored the same number of recordings.

In order to interpret the computed Kappa values, we have used as a reference the interpretation of the Kappa Coefficient provided in (DG, 1991), which attributes good agreement to Kappa values within the interval $[0.61 - 0.80]$ and very good agreement to higher Kappa values $[0.81 - 1.00]$. According to this interpretation Table 4 reveals that: (a) there is good inter-human agreement for both the NAEP-2 and NAEP-4 tasks, (b) there is good machine-to-human agreement for the NAEP-4 task, and (c) there is very good machine-to-human agreement for the NAEP-2 task. It can be observed that the Kappa agreement between the machine and the humans is significantly higher than the agreement between each human scorer and the rest of the human scorers. This is true for both the NAEP-2 and NAEP-4 tasks. This difference in agreement is statistically significant ($z$-test at 5% significance level, $|z| > 1.96$), which shows the ability of the proposed features and classification scheme to provide a useful method to automatically assess expressive oral reading according to the NAEP scale.

It can also be observed in Table 6 a very similar agreement between a human scorer and the rest of human scorers for the NAEP-4 task, which is the most difficult one, this indicates that our humans-scorers were able to administer this rubric quite consistently.

In terms of the Spearman’s rank correlation coefficient ($\rho$) we have found a big inter-human correlation (80 and 81) and an even bigger machine-to-human correlation (86) in the NAEP-4 task. This indicates that NAEP-scores from every pair of scorers are closely related, which is consistent with the weighted Kappa values obtained.

Table 7 shows the Kappa agreement and Spearman’s rank correlation coefficient of scores from reading experts and trained human scorers in the 4-way NAEP scale. Agreement and correlation were computed on the subset of 70 recordings that were scored by both groups of scorers. It can be seen that both the agreement and correlation are higher for the reading experts than for the trained human scorers, this is not surprising since reading experts are more familiar with the rubric and can recognize better the characteristics of fluent reading. Nonetheless, agreement and correlation for the trained humans are high. Finally it can be seen that, while the agreement between both groups of scorers is not high for the 70 recordings, the correlation is still very high. This means that, while the trained human scorers are interpreting the rubric differently than the reading experts, they are still applying it consistently.

### Table 6
Inter-rater agreement and correlation coefficients on the NAEP scale.

<table>
<thead>
<tr>
<th>Scorer</th>
<th># Samp.</th>
<th>NAEP-2 $p(a)$</th>
<th>NAEP-2 $p(e)$</th>
<th>$\kappa$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$</td>
<td>751</td>
<td>0.87</td>
<td>0.50</td>
<td>0.73</td>
<td>0.66</td>
</tr>
<tr>
<td>$H_2$</td>
<td>391</td>
<td>0.90</td>
<td>0.50</td>
<td>0.80</td>
<td>0.69</td>
</tr>
<tr>
<td>$H_3$</td>
<td>698</td>
<td>0.87</td>
<td>0.50</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>$H_4$</td>
<td>799</td>
<td>0.86</td>
<td>0.50</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>$H_5$</td>
<td>367</td>
<td>0.86</td>
<td>0.50</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>$M$</td>
<td>1776</td>
<td>0.94</td>
<td>0.50</td>
<td>0.84</td>
<td>0.77</td>
</tr>
</tbody>
</table>

### Table 7
Kappa agreement and Spearman’s rank correlation coefficient for reading experts and trained human scorers in the 4-way NAEP scale.

<table>
<thead>
<tr>
<th>Scorer</th>
<th>$\kappa$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading experts</td>
<td>0.7243</td>
<td>0.8511</td>
</tr>
<tr>
<td>Trained humans</td>
<td>0.6506</td>
<td>0.7809</td>
</tr>
<tr>
<td>Experts vs trained humans</td>
<td>0.5583</td>
<td>0.7564</td>
</tr>
</tbody>
</table>

5.4. Feature relevance

In this section we provide insight on the relevance of the proposed features to the classification. In particular we
focus on the feature weights resulting from training the classifier that addresses the NAEP-2 classification task (non-fluent versus fluent). For each feature we are interested in the absolute value of the weight, which tells us how relevant the feature is, and also the sign of the weight, which tells us whether the feature is working as intended. The decision function resulting from training this classifier assigns positive values to recordings from non-fluent readers and negative values to those of fluent readers. Thus, features designed to correlate with non-fluent reading (e.g., number of filled pauses) are expected to receive positive weights, while features designed to express fluent reading (e.g., length of word groupings) are expected to receive negative weights. Table 8 summarizes the weights of the proposed lexical and prosodic features.

Looking at lexical features in Table 8, we can see that L1, L4, L2, and L5 have a high weight (in absolute value) and behave as intended in terms of their sign; while L1 and L2 correlate negatively with non-fluent reading, L4 and L5 correlate positively with non-fluent reading. L1 and L2 are, respectively, the WCPM and the speaking rate; we obviously expected them to receive a high weight. L4, which represents the number of trigram back-offs, also has a high weight, back-offs are connected to reading disfluencies like inserting words, skipping words or making decoding errors so L4 is expected to have a significant contribution to the classification.

It is interesting to note the high weight of feature L5, which expresses the variance of the sentence reading rate, this result is supported by the reading literature (Benjamin and Schwanenflugel, 2010). On the other hand, L3, which is the number of repetitions across the reading session, although correlating positively with non-fluent reading, receives a very low weight which means that it is not a very informative feature compared to the others.

Looking at the weights that prosodic features receive we see that P12 (average syllable length) plays a major role and correlates positively with non-fluent reading. According to this result the longer the average length of a syllable the less fluent the reader is. In (Duong et al., 2011) duration features were found to be among the most informative. P5, which is the average number of words between two silence regions and is directly connected to the length of word groupings, has also a high weight in absolute value and works as intended by correlating negatively with non-fluent readers. P6 and P9 are respectively the number of silence regions and the number of filled pauses and, as expected, both correlate positively with non-fluent reading. P10 and P14 also contribute to the classification, although to a lesser degree. P14, which is the only pitch-based feature proposed in this work, helps the classification but only marginally. In particular, when P14 is omitted the classification accuracy on the NAEP-2 task drops from 89.02 to 88.87%. For this reason, and considering the importance that pitch has shown in the automatic assessment of oral reading prosody (Mostow and Duong, 2009; Duong and Mostow, 2010; Duong et al., 2011), we believe that more work is needed in order to identify pitch-based features that can be used independently of the text.

It is interesting to note that P1, which is the ratio of pauses at punctuation marks to the total number of punctuation marks did not behave as expected. These findings are similar to those of Miller and Schwanenflugel (2006) who also found less fluent readers were more punctuation-driven than fluent readers.

The rest of the proposed prosodic features are weighted below 0.25 in absolute value and are not included in the list. Nonetheless they still provide marginal information to the classification since removing them causes a slight deterioration of the classification accuracy. To this effect we note that if only the top 7 lexical and prosodic features are used the classification accuracy on the NAEP-2 task drops from 90.72% to 89.85%.

We believe that there is significant noise in the feature extraction process caused mainly by inaccuracies in the recognition hypotheses. These inaccuracies consist of word-level errors (recall that WER was about 10.7%), inexact word, syllable and phone level alignments, inserted and deleted filled pauses and silences, etc. In addition we noticed that the syllabification tool used in this work did not always produce optimal results. We believe that all of these factors might deteriorate the discriminative power of some features. For this reason weights reflected in Table 8 might not reflect the true potential of some of the proposed features.

### 6. Discussion and conclusions

#### 6.1. Summary of results

Our research investigated fully automatic assessment of expressive reading in children’s oral reading of text passages using a standardized rubric, the 4-point NAEP ORF scale. Lexical and prosodic features were designed to classify one minute recordings of texts read aloud by primary school children on the 4-point fluency scale, and to identify children’s oral reading using the binary classification “not-fluent” (levels 1 or 2) or “fluent” (levels 3 or 4). The proposed system was evaluated based on its classification accuracy, agreement and correlation with respect to scores from trained human scorers.

| Feature | $w_j$ | $|w_j|$ | Feature | $w_j$ | $|w_j|$ |
|---------|-------|-------|---------|-------|-------|
| L1      | −1.56 | 1.56  | P12     | 1.37  | 1.37  |
| L4      | 1.51  | 1.51  | P5      | −0.91 | 0.91  |
| L2      | −1.34 | 1.34  | P6      | 0.82  | 0.82  |
| L5      | 1.33  | 1.33  | P9      | 0.57  | 0.57  |
| L3      | 0.32  | 0.32  | P10     | 0.42  | 0.42  |
|         |       |       | P14     | 0.31  | 0.31  |
|         |       |       | P1      | 0.27  | 0.27  |
| Rest    |       |       |         | < 0.25|       |

Table 8: Feature weights.
Based on judgments from two human judges for each text passage read aloud, the proposed system had an accuracy of 90.93% classifying recordings according to the binary NAEP scale (“fluent” versus “non-fluent”) and 76.05% on the more difficult 4-point NAEP scale (Fig. 1 shows the distribution of recordings across the NAEP fluency levels). For both scales the Kappa agreement between each human scorer and the rest of the human scorers was good, while the Kappa agreement between the machine and the human scorers was good and very good respectively. In addition, the Kappa agreement between the machine and each human scorer was always significantly higher than the Kappa agreement between the human scorers. In terms of the Spearman’s rank correlation coefficient ($\rho$), correlation between the machine and each human scorer was always significantly higher than the correlation between humans.

Based on these results, we conclude that, while human experts are effective at administering the NAEP rubric, the proposed system was found to administer it with higher consistency relative to trained human scorers. In addition, both the lexical and prosodic features proposed, although highly correlated by definition, provided complementary information that led to improved discrimination across fluency levels. Based on the additional gains derived from the use of prosodic features, it appears that these features successfully capture some of the aspects of fluency connected with expression that cannot be captured using lexical features alone.

6.2. Towards valid assessment of expressive oral reading

Automated assessment of children’s reading expressiveness needs, like any other kind of assessment, to demonstrate its validity. To a large extent, since our automated system mimics the behavior of humans rating reading fluency using the NAEP, our validation of our automated system can benefit from prior validation efforts of the NAEP scale itself. The NAEP scale has been used in a variety of research study types (Daane et al., 2005; Kuhn, 2005; Pinnell et al., 1995; Valencia et al., 2010) and found to be a valid indicator of the reading behavior of elementary school children (Daane et al., 2005; Valencia et al., 2010). Further, in all of these studies the NAEP has been found to be a good indicator of reading skills as well as a good indicator of short-term progress in fluency (Kuhn, 2005).

Kane’s argument-based validation framework (Kane, 2006) stresses the importance of making inferences from assessment results based on evidence outlined in interpretive arguments. Throughout this study we have argued that automated assessments of a child’s fluency should resemble those determined by experts and highly experienced teachers. We have argued that this assessment should have similar validity and reliability to teachers and experts in identifying students who need immediate help and for making decisions about reading instruction. We have demonstrated that the system derives ratings that may be more reliable when comparing ratings to members of these groups than members of these groups are among themselves, and that the FLORA system has very high correlation with ratings from these groups. Moreover, FLORA picked out a similar percentage of children as human raters did as being ones deemed as not fluent.

A desirable property of a rater is the ability to produce consistent ratings when evaluating the child’s oral reading of a text upon subsequent encounters with the same oral reading sample (i.e., demonstrate good intra-rater reliability). While human raters might produce slightly different evaluations when presented with the same oral reading sample multiple times, the proposed computer-based rating system is deterministic, it will thus achieve the same score at each encounter with the oral reading sample. In this regard, this approach will be more reliable than human raters.

In addition, in support of validity arguments for automatic expressiveness rating by FLORA, we would argue that automated ratings should be strongly correlated with spectrographic measures of reading prosody related to fluency. Because they are derived from these features directly, we can assume that they are strongly correlated with them.

Finally, we would argue that automated ratings of reading expressiveness should be strongly correlated with other elements of fluency such as rate and accuracy measures. Indeed, our automated ratings of reading expressiveness are significantly informed by scores of WCPM obtained by FLORA (Bolanos et al., 2011).

6.3. Why is it important to automate assessments of oral reading fluency?

Reading assessments provide school districts and teachers with critical and timely information for identifying students who need immediate help, for making decisions about reading instruction, for monitoring individual student’s progress in response to instructional interventions, for comparing different approaches to reading instruction, and for reporting annual outcomes in classrooms, schools, school districts and states. One of the most common tests administered to primary school students is oral reading fluency. Over 25 years of scientifically-based reading research has established that fluency is a critical component of reading and that effective reading programs should include instruction in fluency (Kuhn and Stahl, 2000; Fuchs et al., 2001; Good et al., 2001; Chard et al., 2002; National Reading Panel, 2000). While oral reading fluency does not measure comprehension directly, there is substantial evidence that estimates of oral reading fluency predict future reading performance and correlate strongly with comprehension (Fuchs et al., 2001; editorShinn, 1998). Because oral reading fluency is valid, reliable and relatively easy to administer, it is widely used to screen individuals for reading problems and to measure reading progress over time.
Assessments of oral reading fluency are typically administered to individual students by individual testers, usually the student’s teacher. In a typical administration, the student is presented with a printed grade level text passage that they read for one minute. The teacher uses a stop watch to time the reading, listens to the student reading the passage, and marks those words on a printed copy of the text that the student misreads or skips over. After the student has read one or more passages, the teacher subtracts the number of word errors—i.e., the words she/he has crossed out on the text—from the total words read in the text at the end of one minute to determine the words correct per minute score. This score can then be mapped to WCPM scores for students at their grade level for each trimester of the school year to determine the student’s percentile relative to other students in the US.

Manual administration of these tests requires an inordinate amount of time for individual teachers. In the school district in which we collected the recorded stories used in our study, the director of the literacy curriculum estimated that each first, second and third grade teacher spends about 6 days per year assessing their students’ reading abilities. Automating tests of oral reading fluency, as well as other tests of reading ability (e.g., individual word recognition, vocabulary knowledge and comprehension) would save millions of hours of teachers’ time each year that could be devoted to instruction. Moreover, computer-based tests can provide a lasting record of each student’s recorded speech and performance, and the recordings can be analyzed to identify specific reading problems. There are thus great potential benefits to educators and students of automating oral reading fluency assessment.

To our knowledge, there are currently no fully automated tests of oral reading fluency that incorporate accuracy, rate and expressiveness. Our previous research suggests that fully automatic assessment of WCPM may be feasible (Bolanos et al., 2011). This study suggests that, in addition to estimating WCPM automatically, it may be feasible to also automatically assess how expressively individual students read grade level texts.

A second major benefit of automatic assessment of oral reading fluency is the potential to integrate these estimates into computer-based reading tutors to provide students with immediate feedback about their oral reading behaviors and performance. An obvious application is a computer-based tutoring system that provides students with feedback on the accuracy, rate and expressiveness of each successive reading of the text. The report of the National Reading Panel (2000), which conducted a synthesis of scientifically-based research to identify evidence of effective approaches to reading instruction, identified repeated oral reading of texts, with guidance and feedback, as a highly effective approach to reading instruction. In a computer-based reading tutor, a student is presented with a text at their reading level for timed oral reading. The student’s speech is recorded and processed using algorithms that estimate accuracy and expressiveness. Feedback is presented to the student immediately about their oral reading fluency in terms of their WCPM and expressiveness. Following each reading of the text, those words scored as incorrect are displayed for the student to listen to and practice reading and saying. In addition, students can listen to and practice saying complete sentences that were identified by the system as disfluent or having poor prosody. After practicing these words and sentences, the student is asked to read the text a second or third time, with feedback on their oral reading fluency following each reading, and practice between readings. This example of a computer reading tutor shows how accurate automatic assessment of the three components of oral reading fluency could support individualized instruction in a computer tutor to implement a proven instructional approach for improving and reading and comprehension.

There are significant potential benefits of incorporating measures of expressiveness into assessments of oral reading fluency. One of the major criticisms of using WCPM to measure individual student’s improvements in reading over time, (that is, in response to instruction) is that students strive to read texts as quickly as possible in order to increase their WCPM scores. When a student’s ability is measured in terms of how quickly they can read words in a text in one minute, teachers and students learn to focus on fast reading, at the expense of reading texts expressively, and for comprehension. Thus, using only rate and accuracy as measures of reading ability can undermine the ultimate goal of reading instruction—to help students learn to read words accurately and automatically, at a natural reading rate, and to develop reading strategies that result in deep comprehension of tests (Snow et al., 1998). Incorporating estimates of expressiveness into the overall oral reading fluency measure could alleviate this problem. If how expressively the student reads is incorporated into the oral reading fluency measure, students will be penalized for reading too quickly, since the oral reading fluency score will be lowered when they do not pause between sentences, or produce sentences with appropriate prosody, which requires a slower reading rate in order to realize intonation contours and emphasize new and important information by lengthening specific words. Future research that seeks to understand how measures of accuracy, rate and expressiveness can be combined to optimize reading for comprehension is likely to produce more effective assessments, consistent with the goal of helping students learn to read fluently for good comprehension.

Finally, the realization of accurate automatic assessment of the three components of oral reading fluency provides an unprecedented opportunity for researchers to investigate richer and more informative assessments and instructional approaches that integrate these measures. We expect that future research, and future large scale data collection efforts that measure the three components of oral reading fluency will provide significant new knowledge about how to use this information to assess and improve reading instruction for fluent reading with good comprehension.
6.4. Limitations and need for future work

In sum, this study, when combined with results reported by Bolan˜os et al. (2011), provides preliminary evidence in support of the hypothesis that all three components of oral reading fluency—word recognition accuracy, reading rate, and expressiveness—can be measured with accuracy that approaches human judgments of these measures. We have argued that major benefits to students and teachers can be achieved if future research demonstrates the feasibility of realizing this goal.

A major limitation of these studies is the small number of students that have been used in our research. In order to demonstrate the feasibility of fully automatic assessment of oral reading fluency, speech data during oral reading of leveled texts must be collected for a large and diverse population of students at different grade levels, so that automated estimates can be compared to human ones. Still, the results of our preliminary studies are most promising, and motivate the need for further research.

Acknowledgments

We gratefully acknowledge the help of Angel Stobaugh, Director of Literacy Education at Boulder Valley School District, and the principals and teachers who allowed us to visit their schools and classrooms. We appreciate the amazing efforts of Jennifer Borum, who organized the FLORA data collection effort, and the efforts of Linda Hill, Suzan Heglin and the rest of the human experts who scored the text passages. This work was supported by U.S. Department of Education award number R305B070434, National Science Foundation award number 0733323 and NIH award number R43 DC009926-01.

References


chapter chapter


