FLORA: Fluent Oral Reading Assessment of Children's Speech

DANIEL BOLAÑOS, Boulder Language Technologies and University of Colorado
RONALD A. COLE, Boulder Language Technologies
WAYNE WARD, Boulder Language Technologies and University of Colorado
ERIC BORTS and EDWARD SVIRSKY, Boulder Language Technologies

We present initial results of FLORA, an accessible computer program that uses speech recognition to provide an accurate measure of children's oral reading ability. FLORA presents grade-level text passages to children, who read the passages out loud, and computes the number of words correct per minute (WCPM), a standard measure of oral reading fluency. We describe the main components of the FLORA program, including the system architecture and the speech recognition subsystems. We compare results of FLORA to human scoring on 783 recordings of grade level text passages read aloud by first through fourth grade students in classroom settings. On average, FLORA WCPM scores were within 3 to 4 words of human scorers across students in different grade levels and schools.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*

General Terms: Design, Performance

Additional Key Words and Phrases: Oral reading fluency, fluency assessment, speech recognition, speech verification, reading tracker

ACM Reference Format:
Bolaños, D., Cole, R. A., Ward, W., Borts, E., and Svirsky, E. 2011. FLORA: Fluent oral reading assessment of children's speech. ACM Trans. Speech Lang. Process. 7, 4, Article 16 (August 2011), 19 pages.
DOI = 10.1145/1998384.1998390 http://doi.acm.org/10.1145/1998384.1998390

1. INTRODUCTION

Oral reading fluency is frequently used, along with other measures, to assess an individual's reading level and proficiency. Oral reading fluency is defined as the ability to read text quickly, accurately and with proper expression [National Reading Panel 2000]. Individuals who read text fluently read aloud at a normal speaking rate and with appropriate expression and prosody, as if they were speaking to another person.

Reading assessments provide school districts, teachers, and parents with critical and timely information. This information is used for identifying students who need immediate help, for making decisions about reading instruction, for monitoring the students' progress throughout the school year, for comparing and evaluating reading programs and for reporting annual academic outcomes. Elementary and middle schools throughout the U.S. administer oral reading fluency assessments at the onset of the

Research and Development of FLORA was funded by an NICRD SBIR grant to Mentor Interactive Inc. (R. Cole), by an NSF EHR Grant to Boulder Language Technologies (W. Ward), and by an IES CSL grant to Boulder Language Technologies (W. Ward). The conclusions expressed in this article do not necessary represent the opinions of the funding agencies.
Author's address: email: Daniel.bolanos@colorado.edu.

DOI 10.1145/1998384.1998390 http://doi.acm.org/10.1145/1998384.1998390

school year to screen students for reading problems and periodically thereafter to monitor students' reading progress in response to instruction.

There are significant potential benefits to automating or partially automating the assessment process, 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. In addition, 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.

This article describes the system architecture, technology components, and performance of FLORA, a fully functional Web-based system that estimates individual student's oral reading fluency in a session that takes about five minutes. We evaluated FLORA on approximately 13 hours of speech collected from 313 first through fourth grade students who read grade level text passages. Words correct per minute (WCPM) scores computed by FLORA were compared to scores provided by two independent human judges.

The remainder of the article is organized as follows: In Section 2 we provide the scientific rationale for developing FLORA and summarize previous work on automatic fluency assessment. Section 3 describes FLORA's architecture and implementation. Section 4 describes the speech recognizer and reading tracker. Section 5 describes the data collection and results, and Section 6 presents conclusions and discusses future work that aims to improve performance.

2. BACKGROUND

2.1. The Importance of Assessing Oral Reading Fluency

Based on synthesis of scientifically-based reading research, the National Reading Panel [2000] concluded that "Reading fluency is one of several critical factors necessary for reading comprehension, but it is often neglected in the classroom. If children read out loud with speed, accuracy and proper expression, they are more likely to comprehend and remember the material than if they read with difficulty and in an inefficient way." While the ability to read words in texts accurately, at a natural speaking rate and with appropriate prosody is recognized today as a critical component of reading, this was not always the case. In 1983, Richard Allington wrote: "A lack of fluency in oral reading is often noted as a characteristic of poor readers, but it is seldom treated. Oral fluency rarely appears as an instructional objective in reading skills hierarchies, teachers' manuals, daily lesson plans, individualized education plans, or remedial interventions." In the following decades, a resurgence of research on oral reading fluency has led to new theories and knowledge about the nature of fluency, the cognitive processes that are involved in fluent reading, and to new assessment procedures and interventions designed to help children learn to read fluently and with good comprehension [Levy and Hollingshead 1992; Graf and Masson 1993; Rasinski 1989; Rasinski and Zutell 1990; Rasinski 2000; Samuels 2002; Schreiber 1991, 1980; Kuhn and Stahl 2003; Fuchs et al. 2001; Good et al. 2001; Chard et al. 2002]. This research has established that fluency is a critical component of reading and that effective reading programs should include instruction in fluency.

Fluent reading of text depends upon the ability to recognize words quickly and accurately. Automaticity theory [LaBerge and Samuels 1974; Samuels 1985; Wolf 1999] and related verbal-efficiency accounts of reading [Perfetti 1985] hold that students who have learned to decode printed words automatically are able to devote more attention to comprehending what they are reading. According to the theory, readers who have not achieved automaticity during word recognition must devote significant attention to recognizing words at the expense of devoting this attention to constructing meaning, resulting in slower reading times and weaker comprehension. Support for automaticity and the verbal-efficiency theories of reading is provided by the strong association between the speed of reading words, either in word lists or in context, and measures of reading comprehension. Accurate reading speed is both a strong discriminator of reading ability [Perfetti 1985; Jenkins et al. 2003; Lovett 1987; Rupley et al. 1998], and a strong predictor of later reading proficiency [Lesgold and Resnick 1982; Scarborough 1998; Compton and Carlisle 1994].

While oral reading fluency does not measure comprehension directly, there is substantial evidence that estimates of oral reading fluency predicts future progress and correlates strongly with comprehension [Fuchs et al. 2001; Shinn 1998]. Because oral reading fluency is valid, reliable and relatively easy to administer, it is widely used in schools to screen individuals for reading problems and to measure reading progress over time.

2.2. Using Speech Recognition to Assess and Improve Reading Fluency

There is a history of about two decades of research using speech recognition to assess and improve reading. Seminal research conducted by Jack Mostow and his colleagues in Project Listen at Carnegie Mellon University has demonstrated the effectiveness of speech recognition for improving reading fluency and comprehension for both native and nonnative speakers of English [Mostow et al. 2003; Beck et al. 2004; Poulsen et al. 2007; Reeder et al. 2007]. In an interesting approach to measuring oral reading fluency, Mostow et al. [2003] used an ASR system to measure a student's interword latency, defined as the elapsed time between certain words read aloud by the student that were scored as correctly read by the ASR system. They argue that latency "acts as a microscope to allow us to zoom in on the time the student takes to figure out how to pronounce a word, but does not include the time the student requires to actually say the word." Their model of interword latency produced a correlation of over 0.7 with independent WCPM measures of oral reading fluency using grade level passages.

In the context of Project Tball (Technology Based Assessment of Language and Literacy) at UCLA and USC, Black et al. [2007, 2008] investigated oral reading of 55 isolated words produced by kindergarten, 1st and 2nd grade children with the aim of detecting reading miscues automatically, such as sounding-out, hesitations, whispering, elongated onsets, and question intonations. Black et al. developed a speech recognition system that used specialized grammars to model word-level disfluencies using the subword modeling approach developed by Hagen and Pellom [2005]. Scores produced by the recognition system correlated highly (.91) with fluency judgments provided by human listeners.

Zechner et al. [2009] reported preliminary results of a system for automatic scoring of oral reading fluency in text passages and word lists for middle school students. Pearson correlations between automated and human scores were 0.86 for passages and 0.80 for word lists. Li et al. [2007] describe the system architecture and initial performance of a reading coach that runs on both PCs and hand-held devices. On a test corpus of 105 stories read by 3^{rd}, 4^{th}, and 5^{th} grade children, reading errors were detected about 70% of the time using a speech recognition system trained on children's speech that produced word error rates of about 11%.

A series of studies by Bryan Pellom and Andreas Hagen and their collaborators [Hagen et al. 2004; Hagen and Pellom 2005; Hagen et al. 2007] investigated ways to optimize the Sonic speech recognizer for children's speech. The research resulted in a reduction in the word error rate (WER) from 17.4% to 7.6%. (We note that, while speech recognizers are measured in terms of WER, that is, the number of substitutions, deletions and insertions divided by the total number of words in the reference, insertions are not counted as errors in measures of oral reading fluency, they are simply ignored since they do not affect the number of words read correctly per minute).

Hagen et al. [Hagen et al. 2004; Hagen and Pellom 2005; Hagen et al. 2007] developed a version of Sonic that uses subword modeling. The motivation for this work is the observation that a number of reading disfluencies in children's speech occur at the subword level (e.g., "ba- ba- banana") so they are better modeled using subword lexical units, like syllables, as the basic unit for speech recognition. In the study several subword lexical units and approaches were evaluated for detection of disfluencies and modest gains were reported. Bolaños [2008] reported that additional detection gains can be achieved by using syllable graphs to represent hypotheses from the speech recognition system and to obtain confidence estimates.

Daniel Bolaños [Bolaños 2008] investigated the potential benefits of using Support Vector Machines (SVMs), a powerful Machine Learning classifier approach, to verify children's speech while reading aloud. SVMs were used during a second recognition pass to reclassify subword units previously recognized using Sonic. His research led to further reductions in classification error rates during oral reading when tested on the same children's speech data used in Hagen's [2006] research.

Romanyshyn [2007] compared human and computer scoring of oral reading fluency using Sonic. During a training session college students were trained to score words as read correctly or incorrectly produced in recordings of children reading texts out loud from the test set of the CU Read and Summarized Stories Corpus [Cole and Pellom 2006]. Raters were trained to mark words on a printed copy of the text that were skipped over, mispronounced or substituted for other words. At least two judges scored each recorded passage independently. The judges scored the entire passage (not just one minute), and were able to start and stop the recording while scoring. The WCPM score for each successive minute of speech was compared for the two judges. The WCPM score for Sonic was compared to the scores produced by each judge to produce the average agreement between Sonic and the two judges. The average agreement between the two judges was about 95%. The average agreement between Sonic and each of the judges was 92%.

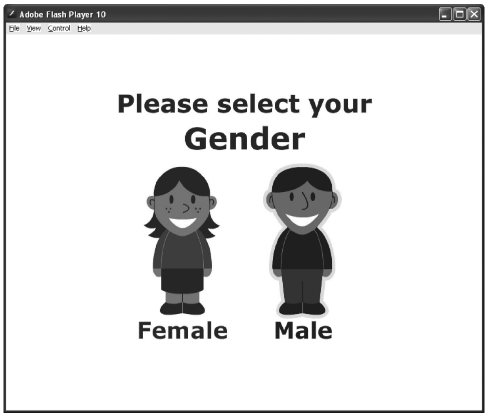
3. THE FLORA SYSTEM

3.1. System Overview

FLORA is a fully functional oral reading assessment system that can be accessed online from any major Webbrowser and used on any PC or Mac. It requires the user to read text passages aloud using a microphone. Speech processing and data persistence are managed by the server machine. FLORA currently runs in two different modes, which reflect the alternative methods of assessing oral reading fluency in schools today. In *cold reading*, the student is presented with a text passage at his or her grade level, and is instructed to read the passage out loud and to skip those words that he or she cannot read. In *reading with word assists* the student is instructed to read the passage out loud, but a teacher says each word that the student is unable to read within three seconds. When FLORA is used in assisted reading mode, the student can use a mouse to click on words they cannot read, and the words are spoken by the system.

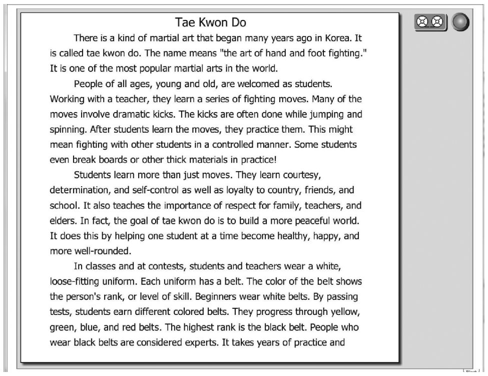
Student Interface. During system use, FLORA supports the following features.

(a) *FLORA* enables an administrator, teacher or student to enroll in the system by providing information about the student's gender, age, and grade level. FLORA



Please select your
Gender

Fig. 1. FLORA screenshot showing the gender selection.



Tae Kwon Do

There is a kind of martial art that began many years ago in Korea. It is called tae kwon do. The name means "the art of hand and foot fighting." It is one of the most popular martial arts in the world.

People of all ages, young and old, are welcomed as students. Working with a teacher, they learn a series of fighting moves. Many of the moves involve dramatic kicks. The kicks are often done while jumping and spinning. After students learn the moves, they practice them. This might mean fighting with other students in a controlled manner. Some students even break boards or other thick materials in practice!

Students learn more than just moves. They learn courtesy, determination, and self-control as well as loyalty to country, friends, and school. It also teaches the importance of respect for family, teachers, and elders. In fact, the goal of tae kwon do is to build a more peaceful world. It does this by helping one student at a time become healthy, happy, and more well-rounded.

In classes and at contests, students and teachers wear a white, loose-fitting uniform. Each uniform has a belt. The color of the belt shows the person's rank, or level of skill. Beginners wear white belts. By passing tests, students earn different colored belts. They progress through yellow, green, blue, and red belts. The highest rank is the black belt. People who wear black belts are considered experts. It takes years of practice and

Fig. 2. FLORA screenshot showing the story to read.

logs the date and time of each session. One of the enrollment screens is shown in Figure 1.

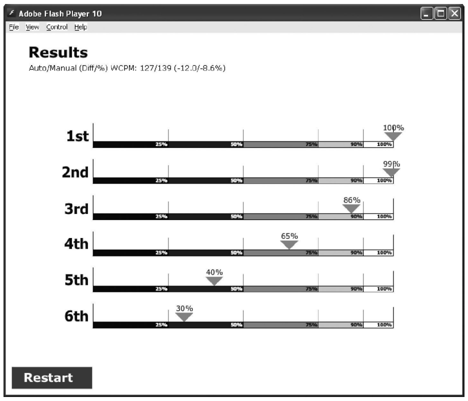
(b) *FLORA* instructs the student that a text will be displayed for reading out loud, and following a countdown (3, 2, 1 . . . Start reading NOW) displays the text. Figure 2 shows the text as it is displayed in the screen for the student to read. The story text completely fits the screen so no scroll is needed, the font size used is: Tahoma 16.

(c) *FLORA* stops the recording after one minute, and thanks the student.

Teachers Interface. The teacher (or researcher) interface was developed to enable individuals to test the system and to score passages. The teacher interface enables the teacher to:

(a) view the WCPM score computed by FLORA for the student. This can be seen in Figure 3, where the green arrows show the same student's WCPM score on each grade-specific fluency scale. Thus for each grade (1^{st} grade, 2^{nd} grade, 3^{rd} grade, etc.) it shows the percentile in which the student's score falls.

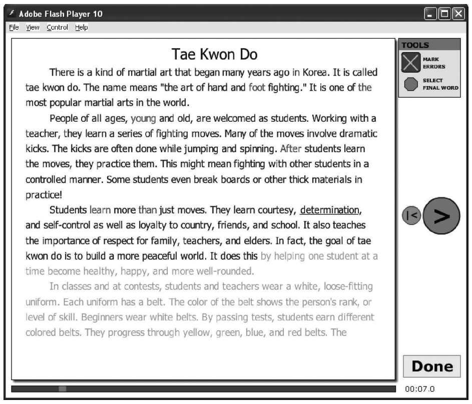
(b) listen to a recording of the student reading a text passage while viewing the text, and click on those words that were skipped or read incorrectly. When the teacher clicks "Done," the WCPM score is displayed, along with the associated percentile for each grade level. The percentile mapping is based on published grade level norms of



Results

Auto/Manual (Diff/%) WCPM: 127/139 (-12.0/-8.6%)

Fig. 3. FLORA screenshot showing the assessment results from a reading session.



Tae Kwon Do

There is a kind of martial art that began many years ago in Korea. It is called tae kwon do. The name means "the art of hand and foot fighting." It is one of the most popular martial arts in the world.

People of all ages, young and old, are welcomed as students. Working with a teacher, they learn a series of fighting moves. Many of the moves involve dramatic kicks. The kicks are often done while jumping and spinning. After students learn the moves, they practice them. This might mean fighting with other students in a controlled manner. Some students even break boards or other thick materials in practice!

Students learn more than just moves. They learn courtesy, determination, and self-control as well as loyalty to country, friends, and school. It also teaches the importance of respect for family, teachers, and elders. In fact, the goal of tae kwon do is to build a more peaceful world. It does this by helping one student at a time become healthy, happy, and more well-rounded.

In classes and at contests, students and teachers wear a white, loose-fitting uniform. Each uniform has a belt. The color of the belt shows the person's rank, or level of skill. Beginners wear white belts. By passing tests, students earn different colored belts. They progress through yellow, green, blue, and red belts. The

Fig. 4. FLORA screenshot showing the support for manual scoring.

oral reading of grade level passages by thousands of students collected during the fall, winter and spring of the school year [Hasbrouck and Tindal 2006]. Figure 4 displays the teacher scoring utility.

3.2. FLORA Architecture and Technology Modules

Figure 5 shows the FLORA architecture, with its modules, data flow and control. The figure also presents information about the communication protocols and the technology utilized.

Client-side. On the clientside, a conventional Webbrowser loads the FLORA Web-application (all major Webbrowsers are supported). The application consists of:

(a) an embedded Flash application that supports all interactions with the user including enrollment, presenting the text, displaying the assessment results, and providing the support for manual scoring.
(b) a Java Applet responsible of the recording of the audio and its transmission to the server by means of a socket. Both the Flash and the Java Applet are synchronized by means of a TCP connection on the client side.

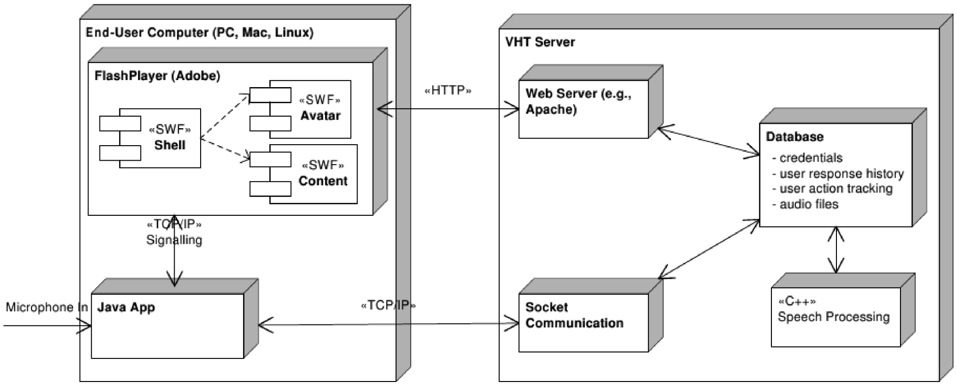


Fig. 5. FLORA architecture: modules and data flow between modules.

Server-side. The server side consists of three modules:

(a) the Webserver.
(b) a Java application responsible for receiving the audio from the client, storing it in an audio repository and sending it to the speech processing module.
(c) the C++ based BLT automatic speech recognizer (BLT ASR) and ReadToMe, the BLT reading tracker. Speech received from the client is recognized in real time so WCPM and percentile scores can be viewed immediately by a teacher (or researcher).

4. SPEECH VERIFICATION DURING ORAL READING

4.1. Speech Recognizer

The speech recognizer used for the development of FLORA is a large vocabulary continuous speech recognition (LVCSR) system written by Daniel Bolaños, supported jointly by BLT and the University of Colorado (CU). Acoustic modeling is based on Hidden Markov Models and Gaussian Mixture Models (HMMs/GMMs).

As in most state of the art systems, speech data were parameterized using Mel-Frequency Cepstral Coefficients (MFCC) and cepstral mean subtraction was applied for noise robustness. Acoustic models were trained under Maximum Likelihood using Baum-Welch reestimation. During the Gaussian-splitting stage, two Gaussian components were added to each mixture after each reestimation 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 specifically designed to match frequent non-speech events.

A trigram language model was trained for each of the stories using the CMU toolkit. The speech recognizer used a static decoding network organized as a prefix-tree and then was compressed using a forward-backward merging of nodes. 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.

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. The regression tree used to cluster the Gaussian distributions comprised 50 base-classes and the minimum occupation count to compute a transform was set to 3500 feature frames. Expectation-Maximization clustering was used to cluster the Gaussian means.

The best results in terms of WCPM were obtained after speaker adaptation, although this only contributed to a relative error reduction of about 7%. Results shown in Section 5.4 are from the adapted system.

Training corpora. Three different speech corpora were used to train the acoustic models used by FLORA. The University of Colorado Read and Summarized Stories Corpus [Cole and Pellom 2006] (325 speakers from 1^{st} to 5^{th} grade), the CU Read and Prompted Children's Corpus [Cole et al. 2006] (663 speakers from Kindergarten through 5^{th} grade) and the OGI Kids' Speech Corpus [Shobaki 2000] (509 speakers from 1^{st} to 5^{th} 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.

4.2. ReadToMe

ReadToMe is BLT's reading tracker, which is built on top of BLT ASR. During an oral reading assessment session, ReadToMe receives audio in realtime from the clientside and returns a WCPM score. The computation of the WCPM score is done as follows.

(1) ReadToMe uses BLT ASR to produce a hypothesis, which is a string of words with a confidence score.
(2) ReadToMe aligns the hypothesis to the reference text (the story read) and tags each of the words in the reference as correctly or incorrectly read (or skipped over). The alignment is done so deletions after the last word correctly read in the reference are not penalized. This makes the hypothesis prone to be aligned to the initial part of the story, which is a reasonable assumption since that is where the student starts to read.
(3) Finally, ReadToMe counts the number of words tagged as correct, which is the WCPM score.

In order to handle speech events that are not words in the reference text, like out-of-vocabulary words and mispronunciations, an all-phoneme ergodic model has been incorporated into the decoding network, this model is intended to match sequences of phones that do not correspond to any pronunciation in the decoding lexicon, it receives a special penalty in order to prevent deletions. An additional set of filler models were trained in order to deal with filled pauses, which are very common in children's speech.

5. DATA COLLECTION AND SCORING

5.1. Data Collection

FLORA was 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. 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 FLORA system 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 either two or three text passages to read aloud.

During the testing procedure, the student was seated before the laptop, and asked to put on a set of Sennheiser headphones with an attached noise-cancelling microphone. The experimenter observed or helped the student enroll in the session that involved (as pictured above) 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, captured the student's speech and relayed the speech to the server.

Table I. Summary of the Data Used for the Evaluation

grade	1^{st}	2^{nd}	3^{rd}	4^{th}	all
# recordings	132	259	165	227	783
# schools	2	3	2	2	4
# students	53	104	66	90	313
hours of audio	2:12′	4:19′	2:45′	3:47′	13:03′

Table II. Statistics of the Stories for Each Grade

grade	*# words*	*# uniquewords*	*# sentences*	*# wordspersentence*
1^{st}	223	110	25	8.92
2^{nd}	249	121	21	11.86
3^{rd}	255	127	20	12.75
4^{th}	381	184	28	13.61

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. Since the goal of our study was to examine FLORA's oral reading fluency scores relative to human judgments for a wide range of students with different reading abilities, the assignment of some students within the same grade to passages at different grade levels does not impact the interpretation of the results. Table I summarizes the FLORA assessment corpus.

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 noncommercial 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 II provides statistics of stories at each grade level.

5.2. Human Scoring of Recorded Stories

In order to evaluate the ability of FLORA to produce reliable WCPM scores, each of the one-minute recordings collected was scored independently by two former elementary school teachers. Each teacher had more than a decade of experience administering reading assessments to elementary school children, which may explain their high overall level of agreement across all recordings. 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.
(2) enabled the scorer to listen to and replay any portion of the recording while viewing the story text.
(3) enabled the scorer to click on each word the scorer judged to be misread or omitted during the one minute recording.
(4) enabled the scorer to click on the last word that was correctly read. This was done so as to compute the total number of words read as accurately as possible. In real world applications, the last word read is determined by the reading tracker.
(5) save the scored text and WCPM score into a database.

Table III. Comprehensive Description of the FLORA Corpus and WCPM Results for Human Scorers and FLORA

				H WCPM		*F WCPM*		*H Diff*		*F to H Diff*	
		# Stud	*#Rec*	μ	(σ)	μ	(σ)	μ	(σ)	μ	(σ)
School 1	*AllGrades*	178	445	78.5	(40.6)	81.0	(40.4)	1.3	(2.61)	3.62	(3.23)
	Grade 1	44	110	36.1	(23.5)	38.0	(23.0)	0.97	(1.89)	2.96	(2.89)
	Grade 2	40	99	73.9	(29.8)	75.9	(29.4)	1.22	(2.31)	3.65	(4.05)
	Grade 3	64	159	92.5	(33.1)	94.7	(32.1)	1.55	(3.31)	3.55	(2.67)
	Grade 4	30	77	116.2	(29.7)	120.0	(29.4)	1.34	(2.12)	4.68	(3.27)
School 2	*AllGrades*	34	86	76.5	(24.8)	78.9	(25.1)	1.19	(1.38)	3.77	(2.87)
	Grade 1	9	22	59.5	(20.9)	62.1	(21.3)	1.77	(1.91)	3.30	(2.26)
	Grade 2	23	58	79.6	(20.9)	81.9	(21.1)	1.02	(1.09)	4.01	(3.11)
	Grade 3	2	6	109.0	(28.0)	111.5	(32.8)	0.67	(0.75)	3.17	(1.93)
School 3	*AllGrades*	41	102	92.8	(44.0)	92.8	(44.0)	1.55	(2.19)	2.91	(3.50)
	Grade 2	41	102	92.8	(44.0)	92.8	(44.0)	1.55	(2.19)	2.91	(3.50)
School 4	*AllGrades*	60	150	130.1	(31.2)	131.4	(29.8)	1.11	(1.14)	4.03	(3.35)
	Grade 4	60	150	130.1	(31.2)	131.4	(29.8)	1.11	(1.14)	4.03	(3.35)
AllSchools		313	783	90.0	(43.0)	92.0	(42.4)	1.28	(2.23)	3.62	(3.27)

The scoring tool was designed to search the corpus of recorded stories and to present unscored stories to each independent scorer until all stories were independently scored by both judges. The average scoring time was about four minutes per story.

5.3. Generation of FLORA WCPM Scores

Each recorded story was processed by FLORA to produce a WCPM score. The score was based on:

(1) gender-independent children's acoustic models,
(2) language models computed for each text passage,
(3) classification of each word in the text as correct, incorrect or skipped over during the minute of speech recorded.
(4) calculation of the WCPM score as the number of words in the text (from the first word in the passage to the final word scored by ReadToMe) minus the number of words tagged as incorrect.

5.4. Results

All results presented below are reported in terms of the WCPM scores provided by the independent human scorers and the corresponding WCPM score produced by FLORA for the same one-minute recording. Because of the high level of agreement and low variance of human scores across all stories, WCPM scores produced by FLORA for each recording are compared to the average of the two human WCPM scores.

Table III presents a summary of all of the results. It shows the number of students and recordings at each grade level in each school, it also shows the mean WCPM scores and standard deviation for all classrooms produced by human scorers and FLORA. Two major results can be seen in this table. Columns 4 (Human WCPM) and 5 (FLORA WCPM) show the mean WCPM scores for each classroom in each school. The numbers in parentheses after each mean WCPM score is the standard deviation from the mean WCPM score. The main result, which can be seen by comparing the adjacent numbers in columns 4 and 5, is that the scores are very similar, as are the standard deviations, for each classroom. This high level of agreement and consistency across classrooms suggests that FLORA provides an accurate measure of WCPM for groups of 20 or more students across classrooms in schools with different student populations and reading performance levels. This result is explored in more detail below.

The second pattern of results is revealed by examining the numbers in column 6, which shows the mean difference in WCPM scores for the two human scores for the recordings in each classroom, and the numbers in column 7, which shows the mean

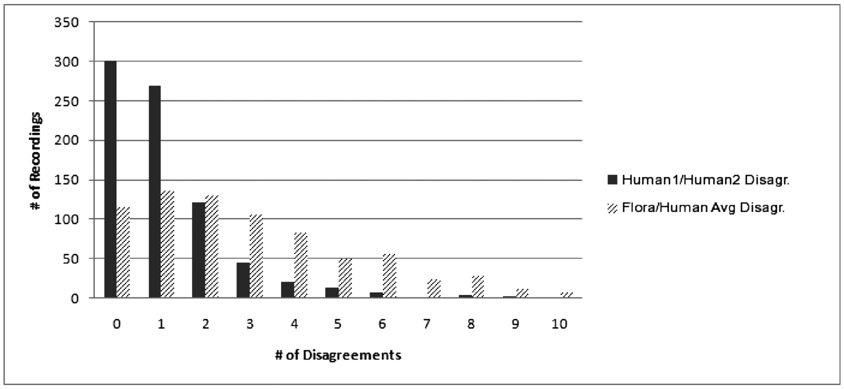


Fig. 6. Distribution of the difference in WCPM between the two human scorers and FLORA and between the two human scorers for the same recordings.

difference between the averaged human scores and FLORA for each classroom. Note that differences in WCPM scores are expressed in absolute value. Viewing the numbers in column 6 reveals the remarkable agreement between the two human scorers (1.28 WCPM difference across all schools) and the low variance. Across all recordings, the mean difference between FLORA and the averaged human scores was 3.62 words, while the mean difference between human scores was 1.28 words. While we are obviously pleased with this number, a comparison of the standard deviations for adjacent numbers in columns 6 and 7 reveals that FLORA scores varied about twice as much as human scores. The nature of this variation is shown in more detail in Figures 6 and 12.

Figure 6 displays the distribution of the difference in WCPM, from 0 to 10, between the two human scorers (dark bars), and between FLORA and the average human score (hatched bars), for all individual recordings. It can be seen that the two human judges produced the same WCPM score for the same story (0 disagreements) 301 times, differed by 1 word 269 times and differed by 2 words 121 times; across all stories human scorers differed by 2 or less words 87.5% of the time and by 4 words or less over 95.8% of the time. Disagreements between FLORA and the average of the two human scores for each story were more evenly distributed, with 562 or 72.7% of scores within 4 words of the averaged human scores.

Figure 7 displays a scatter plot of the WCPM scores from the two human scorers for all recordings, while Figure 8 displays a scatter plot of the WCPM scores from FLORA respect to the average human scores for all recordings. If agreement were perfect, all points would lie on the diagonal.

These figures show the strong agreement between WCPM scores for human scorers on each recording, and the very good agreement between FLORA and human scores, with relatively few outliers.

Figure 9 displays the differences in WCPM scores for humans and FLORA as a function of students' reading rate (based on the averaged human score). It can be seen that at reading rates below 160 WCPM FLORA performs more similarly to human scorers than at higher rates, in which FLORA scores differ from human scores by six or seven words per minute.

It should be noted that students reading at this rate in elementary school are at or above the 90 percentile, so errors are unlikely to affect students.

Estimating WCPM Scores Across Student Populations: The data collection produced a sufficient number of recordings in second grades in three schools to gain some insights

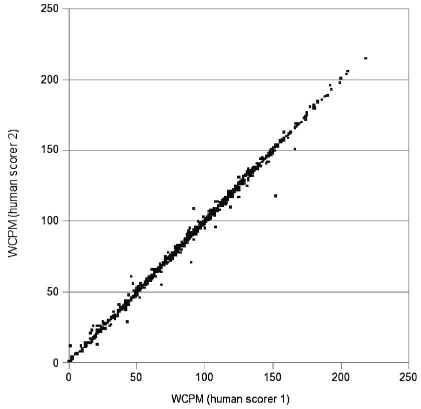


Fig. 7. Correlation between WCPM scores produced by two independent human-scorers on each of the one-minute recordings collected.

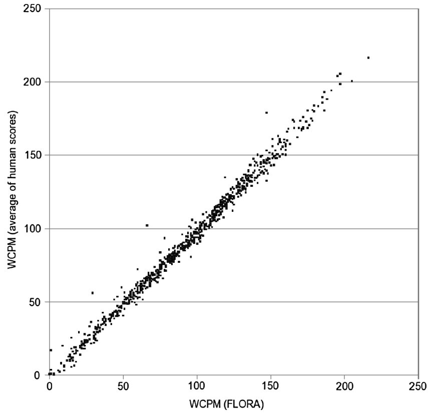


Fig. 8. Correlation between WCPM scores produced by FLORA and the average of two independent human-scorers on each of the one-minute recordings collected.

about FLORA's performance assessing oral reading fluency of students in schools with different levels of student achievement and demographics. 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 language learners 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 English language learners.

Figure 10 displays the mean WCPM scores produced by humans and FLORA for second grade students in each school. As expected, the WCPM scores correlate positively with school literacy performance, with WCPM scores of 75.9, 81.9, and 92.8, respectively, for schools 1, 2, and 3. The main result shown in Figure 10 is that the mean WCPM scores produced for FLORA for each classroom are nearly identical to the human scores. This suggests that FLORA can be used as an accurate estimate of

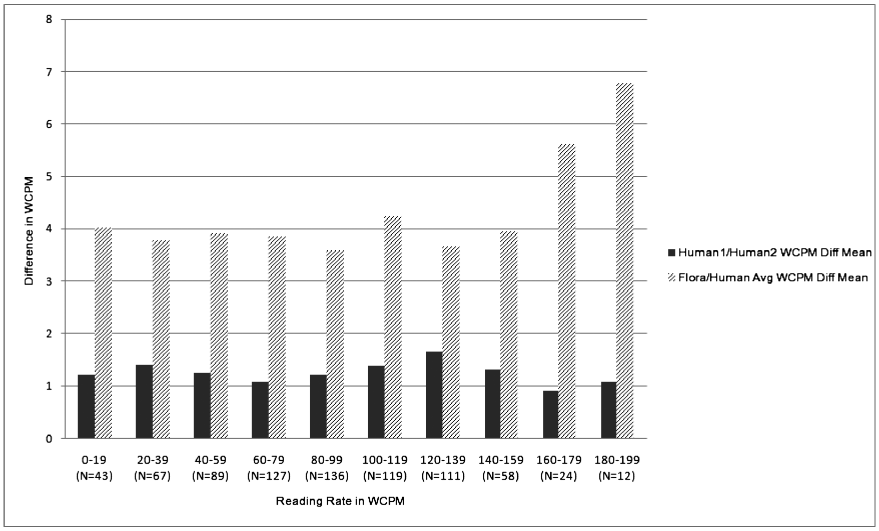


Fig. 9. Differences in WCPM scores for humans and FLORA as a function of students' reading rate.

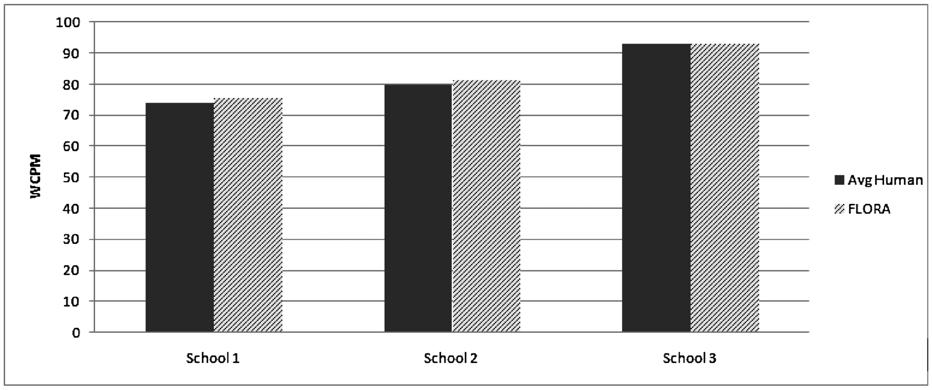


Fig. 10. Mean WCPM scores produced by humans and FLORA for second grade students in each school.

oral reading fluency performance at the classroom level. Figure 11 shows the mean difference in WCPM between humans and between FLORA and the human average for schools 1, 2, and 3. Figure 12 provides further information about FLORA's WCPM performance for second grade students in the three schools by showing the distribution of the difference in WCPM between the two human scorers and between the FLORA WCPM score and the average human score.

In summary, the results indicate that FLORA produces accurate estimates of mean WCPM scores for groups of students in schools with different literacy achievement levels and different ethnographic characteristics. As shown in Figure 12, the distribution of disagreements between FLORA and human scorers is much broader than the distribution of disagreements between human scorers, indicating that human scorers produce more consistently accurate oral reading assessments for individual students than FLORA.

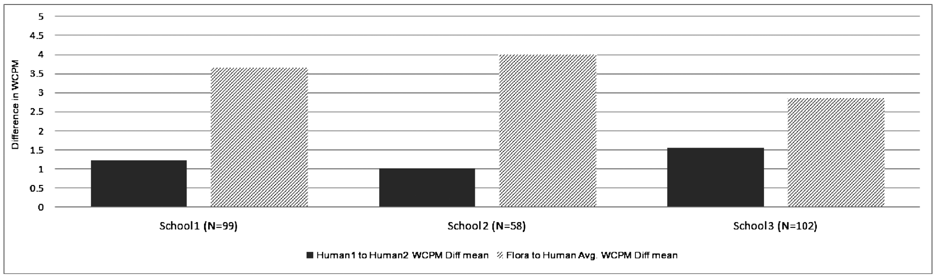


Fig. 11. Differences in WCPM scores for humans and FLORA for second graders across different schools.

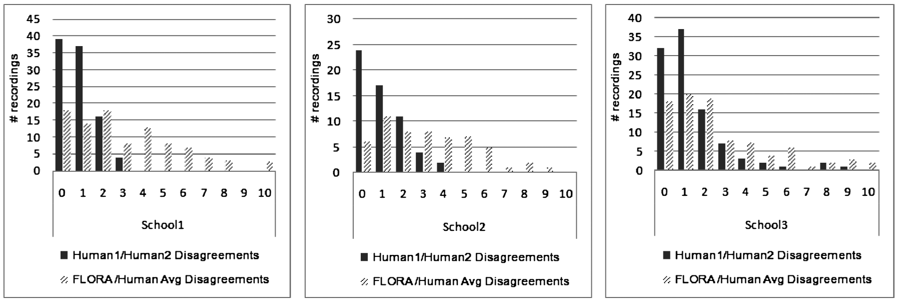


Fig. 12. Distribution of the difference in WCPM between the two human scorers and FLORA and between the two human scorers for the same recordings.

Table IV. Percentiles used for mapping students to percentiles (taken from [Hasbrouck and Tindal 2005])

		percentile				
season	*grade*	$90th$	$75th$	$50th$	$25th$	$10th$
spring	1^{st}	111	82	53	28	15
spring	2^{nd}	142	117	89	61	31
spring	3^{rd}	162	137	107	78	48
spring	4^{th}	180	152	123	98	72

5.5. FLORA as a Tool for Screening Students

At the beginning of each school year, elementary schools across the United States screen entering students to determine if they may be at risk of not learning to read. According to Hasbrouck and Tindal Hasbrouck and Tindal [2006]: “Screening measures help a teacher quickly identify which students are likely “on track” to achieve future success in overall reading competence and which ones may need extra assistance. Screening measures are commonly developed from research examining the capacity of an assessment to predict future, complex performance based on a current, simple measure of performance.” Fuchs et al. [2001] have suggested that oral reading fluency assessments can play a role in screening.

We were interested in determining if FLORA might be a useful tool for providing a WCPM score that could be used as one valuable data point that could be used with other measures to identify at-risk students. One way to do this is to compare human and FLORA WCPM scores to national reading norms developed by Hasbrouck and Tindal [2006]. Table IV summarizes reading norms for the spring semester from 1^{st} to 4^{th} grade taken from these published norms.

Table V shows the relative agreement p(a) for the two independent human scorers and the humans vs FLORA (for each of the human scorers and FLORA) as well as

Table V. Agreement when Mapping Students to Percentiles

	$p(a)$	$p(e)$	κ
$scorer_1$ *vs* $scorer_2$	0.96	0.23	0.95
$scorer_1$ *vs* $FLORA$	0.88	0.24	0.83
$scorer_2$ *vs* $FLORA$	0.87	0.24	0.82

Table VI. Agreement when Deciding Which Students Need Screening

	$p(a)$	$p(e)$	κ
$scorer_1$ *vs* $scorer_2$	0.98	0.60	0.95
$scorer_1$ *vs* $FLORA$	0.96	0.61	0.90
$scorer_2$ *vs* $FLORA$	0.96	0.61	0.90

the Cohen's kappa (κ) in the task of mapping students to percentile ranks. p(e) is the hypothetical probability of chance agreement. As can be seen the relative interhuman agreement is considerably higher than the FLORA to human agreement (0.96 to 0.87). We believe that this difference is due to the higher average difference between FLORA and human WCPM scores, which makes more students in the boundaries to be mapped to the wrong percentile rank. However, we have observed that every time a student is mapped to the wrong percentile by FLORA, it is the nearest percentile to the actual one.

Table VI shows the relative agreement (p(a)) for the two independent human scorers and the humans vs FLORA (for each of the human scorers and FLORA) as long as Cohen's kappa (κ) in the task of identifying students that might need screening. In this task we have used the 50th percentile as cutoff, so all the students that fall below that percentage according to their WCPM score and grade, are tagged as "in need of screening". As can be seen the relative inter-human agreement and the FLORA to human agreement is very close (0.98 to 0.96) and so it is the κ (0.95 to 0.90), which means that FLORA performs very well at identifying students that might require additional reading assessments and instruction.

6. CONCLUSIONS AND FUTURE WORK

We are extremely encouraged by the general pattern of results obtained with the initial FLORA prototype. In the vast majority of recordings, WCPM scores produced by FLORA were close to scores produced by human scorers, with mean differences of 3 to 4 words. The results suggest that FLORA could be used now as a tool for identifying students who may be at risk for learning to read and deserve future attention. It also appears that FLORA may provide an accurate assessment of the mean WCPM produced by a group of students, for example in the classroom, and could thus be used to measure the effectiveness of reading programs or student progress within or across schools at the beginning, middle and end of a school year. In terms of progress monitoring—tracking changes in oral reading fluency through periodic assessment of students in response to instruction—further research is needed to assess the potential of FLORA as a valid and reliable assessment tool. While FLORA provides accurate WCPM scores for the majority of students, it produces more variability for individual students than our expert human scorers. During the next school year, we plan to conduct additional research that will investigate FLORA's performance, relative to human scoring, for the same students over successive months. This research will be conducted using improved speech recognition and reading tracking systems, based on research approaches described in the following.

The potential of FLORA (and similar computer-based oral reading assessments) is obvious and profound. For example, across the U.S., oral reading assessments are administered annually to millions of students. In widely used programs such as DIBELS [Good et al. 2007], these assessments are administered by a teacher to an individual student. The teacher uses a stopwatch to measure one minute of reading, and marks word errors on a sheet of paper that shows the text the student is reading. While this method may be valid and reliable, FLORA provides the added benefit of automating the entire assessment process, providing a digital recording that can be reviewed along with the text (and scored manually by the teacher if desired), enabling the teacher to review the student's speech, make and store notes about the student's reading behaviors, compare the student's current WCPM scores to previous assessments, and listen to recordings of any previous recordings made. These digital records and recordings could be reviewed with parents to help them track and understand their children's reading progress. Moreover, it is easy to imagine extending the capabilities of FLORA, as the system improves through additional research, to create a profile of each student's oral reading behaviors by analyzing the types of reading miscues the student makes and measuring prosodic expression in addition to word accuracy and speed.

There are many obvious ways to improve the current system. The initial prototype, which was completed "just in time" to collect assessment data before the end of the school year, has not been optimized in any way. Our future work will focus on understanding the nature of the errors the system now produces (relative to human judgments), and pursuing established methods of improving the performance of the speech recognizer and reading tracker. These methods will include the following.

—*Acoustic modeling*. Currently the speaker independent version of FLORA uses a single set of acoustic models for all children regardless of gender, age, pitch or language of origin. We plan to use the FLORA speech corpus to train more specific acoustic models. We will also investigate the feasibility of doing semi-supervised training in this scenario in which the reference text is known but not the exact sequence of words that are actually spoken.

—*Speaker adaptation*. We plan to incorporate Vocal Tract Length Normalization (VTLN) as a way to deal with the inter-speaker variability.

—Rejection. We will investigate the utilization of confidence estimates to improve the ability to reject incorrectly read words.

—*Speaker adaptation*. *Text-specific optimization*. Most available oral reading assessments use a specific set of leveled text passages. By collecting and analyzing data from sufficient numbers of children reading these stories using an automated system such as FLORA, it is possible to model common mispronunciations and substitutions, and thus improve performance.

An important goal of our future work is to measure expressive reading automatically. FLORA currently measures reading speed and accuracy, but does not measure how expressively the passage was read. A number of researchers have argued that expressive reading is a critical component of reading fluency, as it indicates that the person comprehends what they are reading. According to Torgesen and Hudson [2006], "One of the most interesting current questions in research on fluent reading concerns the role of prosody in the definition. The role of accuracy and rate seem very central to the notion of fluent reading, but what role does prosody play? Perhaps the most straightforward reason to include prosody as part of the definition of fluency is that it may reflect the reader's understanding of the meaning of the passage being read." Rasinski [2000] argues that this is clearly the case: "When readers embed appropriate volume, tone, emphasis, phrasing, and other elements in oral expression, they are giving evidence of actively interpreting or constructing meaning from the passage. Just as fluent musicians interpret or construct meaning from a musical score through phrasing, emphasis, and variations in tone and volume, fluent readers use cognitive resources to construct meaning through expressive interpretation of the text." While additional research is needed to understand the relationship between expressive reading and comprehension, it is clear that developing valid automatic measures of expressiveness during oral reading will be desired and valued by teachers. Initial research in this area has produced promising results [Mostow and G. Aist 1997; Mostow and Duong 2009], although much work remains to be done in this area, including defining and developing valid and reliable measures of expressive reading.

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 and Suzan Heglin who scored the text passages.

REFERENCES

Beck, J. E., Jia, P., and Mostow, J. 2004. Automatically assessing oral reading fluency in a computer tutor that listens. *Tech. Instr. Cogni. Learn. 2*, 61–81.

Black, M., Tepperman, J., Lee, S., and Narayanan, S. 2008. Estimation of children's reading ability by fusion of automatic pronunciation verification and fluency detection. In *Proceedings of the Interspeech Conference*.

Black, M., Tepperman, J., Lee, S., Price, P., and Narayanan, S. 2007. Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment. In *Proceedings of the Interspeech Conference.*

Bolaños, D. 2008. Advances in the application of support vector machines for continuous automatic speech recognition. Ph.D. thesis, Computer Science Department, Universidad Autonoma de Madrid.

Chard, D. J., Vaughn, S., and Tyler, B. 2002. A synthesis of research on effective interventions for building fluency with elementary students with learning disabilities. *J. Learn. Disabil. 35*, 386–406.

Cole, R., Hosom, P., and Pellom, B. 2006. University of colorado prompted and read children's speech corpus. Tech. rep. TR-CSLR-2006-02, Center for Spoken Language Research, University of Colorado, Boulder.

Cole, R. and Pellom, B. 2006. University of colorado read and summarized stories corpus. Tech. rep. TR-CSLR-2006-03, Center for Spoken Language Research, University of Colorado, Boulder.

Compton, D. L. and Carlisle, J. F. 1994. Speed of word recognition as a distinguishing characteristic of reading disabilities. *Educ. Psych. Rev. 6,* 2, 115–140.

Fuchs, L., Fuchs, D., Hosp, M., and Jenkins, J. 2001. Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scient. Stud. Read. 5*, 239–256.

Good, R., Simmons, D., and Kame'enui, E. 2001. The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scient. Stud. Read. 5*, 257–288.

Good, R. H., Kaminski, R. A., and Dill, S. 2007. Dynamic indicators of basic early literacy skills 6th Ed., Dibels oral reading fluency. http://dibels.uoregon.edu/.

Graf, P. and Masson, M., Eds. 1993. *Implicit Memory: New Directions in Cognition, Development and Neuropsychology.* Lawrence Erlbaum Associates, 49–73.

Hagen, A. 2006. Advances in children's speech recognition with application to interactive literacy tutors. Ph.D. Thesis, University of Colorado at Boulder.

Hagen, A. and Pellom, B. 2005. A multi-layered lexical-tree based recognition of subword speech units. In *Proceedings of the language and Technology Conference (L&TC).*

Hagen, A., Pellom, B., and Cole, R. 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Comm. 49,* 12, 861–873.

Hagen, A., Pellom, B., Van Vuuren, S., and Cole, R. 2004. Advances in children's speech recognition within an interactive literacy tutor. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Langustics—Human Language Technologies*. 25–28.

Hasbrouck, J. and Tindal, G. A. 2005. Oral reading fluency norms grades 1-8. table summarized from behavioral research & teaching. Tech. rep. 33, University of Oregon Behavioral Research and Teaching, http://www.brtprojects.org.

HASBROUCK, J. AND TINDAL, G. A. 2006. Oral reading fluency norms: A valuable assessment tool for reading teachers. *Reading Teach. 59,* 7, 636–644.

JENKINS, J., FUCHS, L., VAN DEN BROEK, P., ESPIN, C., AND DENO, S. 2003. Sources of individual differences in reading comprehension and fluency. *J. Educ. Psych. 95*, 719–729.

KUHN, M. R. AND STAHL, S. A. 2003. Fluency: A review of developmental and remedial practices. *J. Educ. Psych. 95*, 3–21.

LABERGE, D. AND SAMUELS, S. A. 1974. Toward a theory of automatic information processing in reading. *Cogn. Psych. 6*, 293–323.

LESGOLD, A. M. AND RESNICK, L. B. 1982. How reading disabilities develop: Perspectives from longitudinal study. In *Proceedings of the International Conference on Theory and Research in Learning Disability.* Plenum, NewYork.

LEVY, B.A., D. P. R. AND HOLLINGSHEAD, A. 1992. Fluent rereading: repetition, automaticity and discrepancy. *J. Experi. Psych. Learn. Memo. Cogn. 18*, 957–971.

LI, X., JU, Y. C., DENG, L., AND ACERO, A. 2007. Efficient and robust language modeling in an automatic children's reading tutor system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

LOVETT, M. 1987. A developmental approach to reading disability: Accuracy, and speed criteria of normal and deficient reading skill. *Child Devel. 58*, 234–260.

MOSTOW, J., AIST, G., BURKHEAD, P., CORBETT, A., CUNEO, A., EITELMAN, S., HUANG, C., JUNKER, B., SKLAR, M. B., AND TOBIN, B. 2003. Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *J. Educ. Comput. Resea. 29,* 1, 61–117.

MOSTOW, J. AND DUONG, M. 2009. Automated assessment of oral reading prosody. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED'09)*. 189–196.

MOSTOW, J. AND G. AIST, G. 1997. The sounds of silence: Towards automated evaluation of student learning in a reading tutor that listens. In *Proceedings of the 14th National Conference on Artificial Intelligence*. American Association for Artificial Intelligence. Providence, 355–361.

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. Tech. rep.

PERFETTI, C. 1985. *Reading ability*. Oxford University Press, Oxford, UK.

POULSEN, R., WIEMER-HASTINGS, P., AND ALLBRITTON, D. 2007. Tutoring bilingual students with an automated reading tutor that listens. *J. Educ. Comput. Resear. 36,* 2, 191–221.

RASINSKI, T. V. 1989. Fluency for everyone: Incorporating fluency instruction in the classroom. *Read. Teach. 42*, 690–693.

RASINSKI, T. V. 2000. Speed does matter in reading. *Read. Teach. 54*, 146–151.

RASINSKI, T. V. AND ZUTELL, J. B. 1990. Making a place for fluency instruction in the regular reading curriculum. *Read. Resear. Instr. 25*, 85–91.

REEDER, K., SHAPIRO, J., AND WAKEFIELD, J. 2007. The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for canadian immigrant children. In *Proceedings of the 15th European Conference on Reading*.

ROMANYSHYN, N. 2007. Automatic measures of oral reading. M.S. thesis, Computer Science Department, University of Colorado, Boulder.

RUPLEY, W. H., WILLSON, V. L., AND NICHOLS, W. D. 1998. Exploration of the developmental components contributing to elementary school childrens reading comprehension. *Scient. Stud. Read. 2*, 143–158.

SAMUELS, S. J. 2002. Reading fluency: Its development and assessment. In *What Research Has to Say about Reading Instruction*, A. E. Farstrup and S. J. Samuels Eds., International Reading Association, 166–183.

SAMUELS, J. 1985. *Automaticity and Repeated Reading*. Lexington Books, Lexington, MA.

SCARBOROUGH, H. S. 1998. Early identification of children at risk for reading difficulties: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, and A. J. Capute Eds., *Specific Reading Disability: A View of the Spectrum*. York Press, 75–199.

SCHREIBER, P. A. 1980. On the acquisition of reading fluency. *J. Read. Behav. 12*, 177–186.

SCHREIBER, P. A. 1991. Understanding prosody's role in reading acquisition. *Theory Prac. 30*, 158–164.

SHINN, M., Ed. 1998. *Advanced Applications of Curriculum-Based Measurement*. Guilford, New York.

SHOBAKI, K., H. J. C. R. 2000. The ogi kids speech corpus and recognizers. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. ISCA.

TORGESEN, J. AND HUDSON, R. 2006. Reading fluency: critical issues for struggling readers. In *Reading Fluency: The Forgotten Dimension of Reading Success*, S. Samuels and A. Farstrup, Eds., International Reading Association, 166–183.

WOLF, M. 1999. What time may tell: towards a new conceptualization of developmental dyslexia. *Annals Dyslex. 49*, 3–28.

ZECHNER, K., SABATINI, J., AND CHEN, L. 2009. Automatic scoring of children's read-aloud text passages and word lists. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*.

Received June 2010; revised October 2010; accepted January 2011