Using Automatic Speech Recognition to Assess the Reading Proficiency of a Diverse Sample of Middle School Students

Klaus Zechner, Keelan Evanini, Cara Laitusis

Educational Testing Service, Princeton, NJ, USA

kzechner,kevanini,claitusis@ets.org

Abstract

This paper describes a study exploring automated assessment of reading proficiency, in terms of oral reading and reading comprehension, for a middle school population including students with reading disabilities and low reading proficiency, utilizing automatic speech recognition technology. We build statistical models using features related to fluency, pronunciation, and reading accuracy to predict three dependent variables: two are related to accuracy and speed of reading, the third is a reading comprehension measure from a state assessment of reading. The correlation coefficients of the best-performing linear regression models range from $r = 0.64$ (reading comprehension score) to 0.98 (correctly read words per minute). We further look at the features with the highest absolute regression weights in the three models and find that most of them fall into the classes of reading accuracy and reading speed. Still, features from the pronunciation class and other fluency features, e.g., relating to silences in the read speech, are also represented in the regression models but with less emphasis.

Index Terms: reading proficiency, students with disabilities, automated assessment, reading comprehension

1. Introduction

Even though teaching students to read has long been a central part of elementary school education, the more widespread use of comprehensive accountability assessments in the U.S. has drawn more attention to certain subgroups of students, including students with disabilities and English language learners. Since the passage of the No Child Left Behind Act, students with disabilities have been participating in statewide standardized assessments of reading and mathematics in greater numbers. This participation has exposed serious weaknesses in some students' reading proficiency, not only at the end of elementary school, but even throughout middle school and beyond [1].

In addition, research on students with learning disabilities indicates that many students are performing at chance level on state assessments, which results in a frustrating test taking experience. One approach taken by some states is to allow students with reading-based learning disabilities to participate in state assessments with an audio (read aloud) accommodation. While this accommodation allows students to demonstrate their comprehension skills, it removes the constructs[1] of decoding and reading fluency from the assessment. One possible solution to this would be to include a direct measure of oral reading fluency in addition to a measure of listening comprehension. The primary drawback of such a design for large scale assessments is the manual scoring of measures of oral reading fluency which is a costly and time consuming process.

Furthermore, oral reading fluency measures have recently gained more widespread use as one of several tools for progress monitoring in a new process of identifying students with learning disabilities called Response to Intervention (RTI) [2]. One of the most common methods used in RTI is the monitoring of oral reading fluency in struggling readers. If such screening can be done with automated methods, thus saving the time- and labor-intensive human evaluation of reading assessments, more students could benefit from an earlier diagnosis of reading problems. This would result in more students receiving the appropriate remedial instruction sooner than is currently the case.

There have been several studies which indicate that there is a high correlation between traditional reading instruments relating to reading comprehension and oral fluency assessments where speed and accuracy of students' read speech are measured, [3, 4, 5]. Correlations typically fall in the 0.65-0.7 range for predicting untimed passage reading comprehension test outcomes [6]. This result motivates further research into using automated measures of oral reading fluency for a general reading proficiency assessment.

This study uses an approach focusing on predicting measures of oral reading proficiency, as well as reading comprehension, utilizing features derived from an automatic speech recognition (ASR) system. Statistical models are built to predict two reading proficiency measures computed from human annotations of reading errors and

[1] A construct is the set of knowledge, skills, and abilities measured by a test.

a measure of reading comprehension from a state reading assessment.

We use data from U.S. middle school students divided into three sub-groups: students with reading disabilities, students with low reading proficiency, and students with no reading disability. One particularly novel contribution of this paper, aside from looking at a population of middle school students that has a very heterogeneous distribution of reading proficiency, is that we explore the extent to which features from an ASR system that account for three major dimensions of reading proficiency (fluency, pronunciation, and accuracy), can be used to predict two distinct categories of reading scores: (a) scores relating directly to oral reading proficiency ("correctly read words per minute", and "relative number of correctly read words"), and (b) a measure for reading comprehension obtained from a state reading assessment.

Our two main research questions are: (1) How accurately can we predict the three aforementioned measures of reading proficiency using ASR technology? Based on previous research, we conjecture that oral fluency measures, in particular the "correctly read words per minute" measure (which includes a reading time component), will be easier to predict than the measure of reading comprehension. (2) Which features are most correlated with the three measures in question; and, related to this, which features dominate in models predicting the three measures?

The remainder of this paper is organized as follows: Section 2 summarizes related work, both in terms of reading assessment of students in general, as well as related to automated scoring of reading proficiency. Section 3 describes the ASR system, the data we use in our study and how it was annotated by human experts, Section 4 introduces the automated measures we compute for reading proficiency, followed by Section 5 detailing the results. In Section 6, we discuss our findings, before we conclude in Section 7.

2. Related Work

2.1. Traditional assessment of children's reading proficiency

Students take many types of reading assessment such as end of unit classroom assessments, formative assessments, criterion referenced state achievements tests in English language arts, and group administered norm-referenced reading assessments. In addition to these reading assessments, students with learning disabilities participate in norm-referenced individually administered achievement assessments and sometimes formative curriculum based measurement tools. Norm-referenced achievement tests generally include several reading subtests that focus on a specific area of reading (e.g., decoding, reading fluency, vocabulary knowledge, and comprehension) and are generally administered by a school psychologist when a student is evaluated for special education and then on an interim basis to determine if the student still qualifies for receiving special education services. Two of the most commonly administered assessments of this type are the Woodcock Johnson Test of Achievement-III (WJ-III) [7] and the Wechsler Individual Achievement Test Second Edition (WIAT-II) [8].

In addition, it is becoming more common for students with reading-based learning disabilities (as well as struggling readers) to participate in progress monitoring assessments such as Curriculum-Based Measurement (CBM) [9]. CBM generally includes short but frequent (e.g. weekly) assessments of a student's achievement in a specific domain. Common CBM probes for reading have included measures of oral reading fluency (e.g., words per minute, words correct per minute, and percentages of words read correctly per minute) as well as measures for reading comprehension.

2.2. Automated assessment of children's reading proficiency

The automated assessment of read speech produced by adult non-native speakers has been a widely researched topic for several years. Systems have been built using measures related to pronunciation, fluency, and reading accuracy, and have achieved high correlations with human judgments of English proficiency, e.g. [10, 11, 12].

The automated assessment of children's speech is different than the assessment of adults' speech in some ways which make the task more difficult. For example, children's speech exhibits higher fundamental frequencies (F0) than adults on average. Also, children's more limited knowledge of vocabulary and grammar results in more errors when reading printed text. Therefore, to achieve high-quality recognition on children's speech, modifications have to be made on recognizers that otherwise work well for adults.

Following the seminal paper on Project LISTEN [13], a number of systems have been built that attempt to overcome these difficulties and use automatic speech recognition technology to assess children's read speech. According to its main purpose of developing an interactive reading tutor that provides feedback about the correctness of a child's reading, much of Project LISTEN's early work focused on using ASR technology to detect reading miscues [13, 14]. Such a classification task is hard in that the children's speaking deviations from the text may include arbitrary words and non-words. To overcome this, they modified the recognizer's lexicon and language model to model expected variations produced by children. More recently, the project has incorporated a wide array of features relating to pronunciation and prosody, and has built regression models to predict students' fluency and comprehension scores [15, 16].

The Technology Based Assessment of Language and Literacy project (TBALL) [17] is another long-term project that has been attempting to evaluate the language and literacy skills of young children automatically. In the TBALL project, a variety of tests including word verification, syllable blending, letter naming, and reading comprehension are used in combination. Word verification is an assessment that measures the child's pronunciation of read-aloud target words. A traditional pronunciation verification method based on log-likelihoods from Hidden Markov Models is used initially [18]. Then an improvement based on a Bayesian network classifier [19] is employed to handle complicated errors such as pronunciation variations and other reading mistakes. In addition, the project has also conducted research in predicting reading comprehension scores [20].

In another related study, [21] explored to what extent measures of oral reading proficiency, such as "correctly read words per minute", can be obtained by using ASR technology, using a corpus of text passages and word lists read by children. They found that despite a word accuracy of only 72%, correlations of 0.86 between human rater scores and automated scores could be achieved for read-aloud passages. However, their study did not use any ASR features for these predictions, aside from the ASR hypothesis itself, as well as the overall speaking time.

The current study incorporates many of the features and approaches that have been shown to be effective in previous studies for assessing children's reading proficiency, and applies them to a novel domain: the assessment of children with reading disabilities.

3. Speech Recognizer and Data

3.1. ASR training

We train a gender-independent, triphone HMM broadband speech recognizer, using data from the CSLU Kids' Speech corpus [22] and the CMU Kids Corpus [23], as well as in-house data from middle school students' read-aloud passages collected in 2007 to build the acoustic model (AM).

For the language model (LM), we use 494 read passages, based on four different texts, collected from middle school students from a 2009 pilot test that uses the same four texts that are used for the final assessment in 2010 (described below).[2]

There is no speaker overlap between the data used for AM and LM training and the evaluation data described in the following subsection. The average ASR word accuracy across all 547 passages in our data set is 0.80; it ranges from 0.71 to 0.85 across three groups of students with different proficiency levels (see Table 2).

[2]Passages with more than 10% reading errors were excluded from LM building to allow a bias for more well-formed passages.

3.2. Reading data

For our study, we use a total of 547 read-aloud passages, based on four unique texts, from 167 U.S. eighth-grade middle school students[3] who also completed seven reading comprehension passages from the standardized end-of-year state assessment given to all eighth grade students in the state, henceforth referred to as MSSA ("middle school state assessment").[4] The four read-aloud passages in this study were drawn from released forms of their state assessment and then pilot tested before being assembled into a complete test form. The passages range in length from 339 to 372 words, with an average length of 352 words. Two of the texts are of a literary genre, the other two are informational texts.

Table 1 shows the number of students in the three categories "no reading disability" (ND), "low reading proficiency" (LP), and "reading disability" (RD), as well as the means and standard deviations for their MSSA scores[5] and number of completed passages.

Category	# of test takers	MSSA score	# of completed reading passages
No reading disability (ND)	96	246.1 (10.3)	3.7 (0.7)
Low reading proficiency (LP)	9	227.2 (4.7)	3.0 (0.9)
Reading disability (RD)	62	224.8 (10.5)	2.6 (1.3)
Total	167	237.0 (14.6)	3.3 (1.1)

Table 1: Mean (and std. dev.) reading comprehension scores and average number of read passages for each of three groups of students

Students in the reading disability (RD) category were receiving special education services in reading and had been diagnosed as having a Specific Learning Disability by a special education team at the school. Students in the low reading proficiency (LP) group were not receiving services for special education but had performed in the lowest proficiency category on their last state assessment. Finally, the students in the no reading disability (ND) group were randomly selected from a class roster. None of the students in any of the three groups were receiving English as a Second Language (ESL) instruction

[3]The participants' ages were not recorded directly; however, given the fact that all students were in eighth grade, we can assume that the participants are all approximately 13 years of age.

[4]Not all speakers read all four passages, and some passages were not read completely and were excluded from this study, as described in Section 3.3.

[5]The MSSA score is a scaled score ranging from 200-300.

or classified as English Language Learners (ELLs).

3.3. Reading time and error annotation

For practical purposes, each student was given a maximum of four minutes to read each passage. However, several participants were not able to finish reading some or all of the passages due to their slow reading speed and/or reading disability. In this study, we only use the set of 547 completed passages since (1) this increases the comparability across passages; (2) the automatic detection of the last word a student attempted to read is difficult, particularly due to ASR errors; and (3) an incomplete passage by itself is a very strong predictor of very low reading proficiency (87% of incompletely read passages were from low proficiency students and students with a reading disability).

After the data collection, expert human annotators listened carefully to the recordings of the read passages and marked up all errors of word deletions, substitutions and insertions. These errors were then automatically extracted from these annotations.

Table 2 shows the average passage reading time, the absolute passage errors, and the average word accuracy of the ASR system for the three groups of students.

Category	Avg reading time (sec)	Avg reading errors	Avg word accuracy of the ASR system
ND	146.6 (28.7)	13.7 (12.6)	0.85 (0.11)
LP	176.2 (30.7)	30.6 (20.5)	0.74 (0.12)
RD	194.7 (35.0)	42.6 (33.0)	0.71 (0.14)
Total	162.2 (37.8)	23.0 (24.8)	0.80 (0.14)

Table 2: Mean (and std. dev.) passage reading times, passage errors, and word accuracy of the ASR system for three groups of students

4. Measures and Features

4.1. Measures

We use three measures for reading proficiency (dependent variables in our study). The first two are directly related to oral reading proficiency and have been widely used in past research and reading assessments graded by human raters: (1) correctly read words per minute (CWPM) and (2) relative number of correctly read words (REL_CW). The third measure (3) is the aforementioned MSSA score, measuring reading comprehension.

CWPM is a widely used measure for oral reading proficiency and is computed by subtracting deletion and substitution errors from the reference text and then dividing the result by the reading time in minutes. (Insertions are not considered since they increase reading time and are hence penalized already by the time measure, which considers the duration between the first and last words read by the speaker after leading and trailing silences are removed.) REL_CW uses the same formula but is not normalized by time but by the passage length, i.e., is not indicative of reading speed and measures reading accuracy only. Word deletion and substitution counts to obtain these two measures are derived from human expert annotations of the students' reading passages, as mentioned above.

Even though the MSSA score measures reading comprehension and not oral reading proficiency, past studies have shown that these two measures are significantly correlated and one purpose of this study is to verify these past findings on our data, along with investigating which features from the ASR system are the best predictors for the MSSA score. Table 3 looks at the first question, namely the inter-correlations of the three measures we use in this study.

Pair of measures	Inter-correlation
CWPM REL_CW	0.608
CWPM MSSA	0.640
REL_CW MSSA	0.496

Table 3: Inter-correlations between CWPM, REL_CW and MSSA scores

We can observe that the correlation between CWPM and MSSA scores of 0.64 is well in line with findings from previous research that report correlations between oral reading fluency measures and comprehension scores of around 0.65-0.70 [6]. Not taking reading speed into account produces lower correlations ($r = 0.5$, for REL_CW vs. MSSA). Although REL_CW and CWPM are similar measures, they still differ in that only the latter includes reading time; this explains the comparatively lower inter-correlation between these measures (0.61).

4.2. Features

Based on the ASR hypotheses (words and timing information), as well as scores from the acoustic and language models of the ASR system, a set of 18 features was generated, covering the reading proficiency aspects of fluency, pronunciation, and reading accuracy. The accuracy features 1, 2, 3 and 18 in Table 4 were calculated based on the string edit distance between the ASR word hypothesis and the reference passage. Table 4 lists all 18 features, the category they belong to, as well as their correlations with the three measures, CWPM, REL_CW and MSSA scores.

Feature number	Feature class (proficiency aspect)	Description	Corr. with CWPM	Corr. with REL_CW	Corr. with MSSA
1	Accuracy	Estimated absolute student word errors	-0.452	-0.679	-0.444
2	Accuracy	Estimated reading accuracy	0.556	0.679	0.490
3	Accuracy	Estimated REL_CW	0.456	0.683	0.443
4	Accuracy	Normalized LM score	-0.461	-0.406	-0.319
5	Pronunciation and Accuracy	Normalized ASR confidence score	0.601	0.604	0.519
6	Pronunciation	Normalized AM score	-0.444	-0.304	-0.333
7	Fluency	Rate of long silences	-0.467	-0.171	-0.284
8	Fluency	Mean deviation of long silences	-0.364	-0.161	-0.206
9	Fluency	Rate of repetitions	-0.550	-0.439	-0.480
10	Fluency	Total reading time	-0.905	-0.411	-0.585
11	Fluency	Rate of silences (per time unit)	-0.324	*n.s.*	-0.181
12	Fluency	Rate of silences (relative to words spoken)	-0.534	-0.165	-0.325
13	Fluency	Mean of silence duration	-0.307	-0.117	-0.189
14	Fluency	Mean deviation of silences	-0.319	-0.146	-0.204
15	Fluency	Length of uninterrupted phrases (no intervening silences)	0.377	0.122	0.209
16	Fluency	Mean deviation of uninterrupted phrases	0.333	*n.s.*	0.194
17	Fluency	Speaking rate	0.936	0.557	0.622
18	Fluency and Accuracy	Estimated CWPM	0.920	0.629	0.645

Table 4: Description of 18 ASR features, along with their category and correlations with three proficiency measures. (All correlations are significant at $p < 0.05$ except for two fluency feature correlations with REL_CW)

5. Results

We use the machine learning toolkit Weka [24] to investigate several machine learning approaches to predicting the three continuous dependent variables, CWPM, REL_CW, and MSSA scores, based on the 18 features listed in Table 4. (As mentioned above, both CWPM and REL_CW measures are derived from human annotations of the read-aloud passages.) The 18 independent variables were all z-score normalized before being used as features in the machine learning models.

The data set has 547 instances (passages read out loud), and ten-fold cross-validation was used due to the limited size of the data set. One of the three independent variables, the MSSA score, is identical across all instances from a given speaker. However, due to the fact that different speakers have varying numbers of completed reading passages in this data set (as shown in Table 1), no speaker-level aggregation of features from the different reading passages was conducted for predicting the speaker-level MSSA scores (for the same reason, the cross-validation sets were sampled on the response-level, not on the speaker-level). Thus, the speaker-level MSSA scores were predicted for each response from a given speaker, in the same way as the response-level CWPM and REL_CW independent variables were predicted.

Table 5 presents the results from the four highest performing machine learning models based on the correlation between the model predictions and the actual values; the default model parameters provided by Weka were used, except where indicated otherwise in the table.

Classifier	CWPM	REL_CW	MSSA
multilayer perceptron	0.974	0.607	0.490
SVM (linear kernel, epsilon-SVR)	0.978	0.641	0.638
M5 regression tree	0.976	0.702	0.644
linear regression (greedy attribute selection)	0.977	0.710	0.643

Table 5: Correlations between model predictions and actual values for 4 different machine learning approaches

As Table 5 shows, regression-based models consistently had the highest performance across the three data

sets. In order to provide a more detailed analysis of the performance of the linear regression models, Table 6 presents the means of the three dependent variables (CWPM, REL_CW, MSSA), the models' root mean square errors (RMSE), their root relative squared errors (RRSE), the Pearson correlations between the model predictions and actual values, and a list of the features receiving the highest weights in the regression models.[6]

Finally, Table 7 lists the 18 features used in the 3 regression models, sorted by the number of models in which they were selected. The table also presents the parameters of the three linear regression models that were trained for the three data sets; the weights for the features that were selected are included in each column in the table for the three models as well as the intercept (α) of each model. Features that were not selected in a given model are indicated by '–' in Table 7.

6. Discussion

This paper explores the automated scoring of reading proficiency of a sample of middle school students with a very diverse distribution of reading proficiency (students with reading disabilities, with low reading proficiency, and without reading disability). As Tables 1 and 2 show, the group of low proficiency students perform somewhat better than those with reading disabilities, but the gap between the former group and the non-reading-disability group is markedly wider. This shows on all measures: the MSSA scores for reading comprehension, the average number of completed passages, the average passage reading time, and the average rate of reading errors per passage. It is interesting, though, that the smallest relative difference between the low proficiency and the reading disability group is observed for the MSSA scores that measure comprehension. This potentially indicates that features related to oral fluency may be able to distinguish these two groups better than reading comprehension tests, although a larger sample size in the LP group would be necessary to demonstrate this more conclusively.

The two main research questions addressed with this study were (1) to investigate how well features derived from an ASR system can be used to predict scores of both oral reading proficiency as well as reading comprehension for a population of middle school students with heterogeneous reading proficiency; and (2) to look at the relative contributions of these ASR features to the prediction models.

As for the first question, we find that, as initially conjectured, the CWPM measure can be most accurately predicted by automated means; the correlation of the linear regression model is 0.98. The model correlation of the somewhat related oral proficiency measure REL_CW, indicating the relative number of correctly read words in a passage, while being statistically significant, is comparatively much lower (0.71). The main reason is the much lower correlation of the speaking rate feature to REL_CW, compared to CWPM (see Table 4, 0.56 vs. 0.94). As a comparison, [25] describe a study with adult readers in which a measure of correctly read words is predicted (the measure used in that study also included insertions, so it was slightly different from the measure used in the current study). They obtain correlations close to 1.0 with human error annotations; however, it is known that ASR word accuracy for adults can reach 90-95%, whereas in our study involving middle school age children, the ASR word accuracy range is much lower: 71-85%, depending on the students' reading proficiency. As a consequence, estimations of reading errors are much less accurate here for children's speech, compared to the results from [25].

Finally, the regression model correlation with the MSSA reading comprehension score is 0.64, at the same level as the inter-correlation between the human CWPM measure and the MSSA scores.

Interestingly, for all three regression models, the dominant features are from the accuracy class as well as related to overall reading time and speaking rate (Table 7). This finding confirms observations in other contexts of low entropy speech scoring, where features related to accuracy perform significantly better than those related to fluency and pronunciation aspects of speech.

Still, when looking at the entire set of 18 ASR features used for building the three regression models, we find that only two features (relating to properties of uninterrupted phrases) are never selected by the regression models (see Table 7). On the other hand, from the four features selected by all three models, three are related to the internal workings of the ASR system, namely the acoustic model (indicating correctness of pronunciation), the language model and a combination of these (confidence score). Furthermore, seven of nine features selected by two of three regression models are related to fluency, e.g., properties of silences, repetitions etc.

From this, we can conclude that many features related to pronunciation and fluency also play a role in the prediction of the three measures of reading proficiency, but their role is less prominent than that of features related to accuracy and speaking rate.

In addition, beyond the task of predicting a human score of reading proficiency, many features derived from the ASR system may be used as additional feedback to both students and teachers, as well as reading tutors, on more specific strengths and weaknesses of the students' oral reading proficiency.

In summary, given that such automated measures of oral reading proficiency are relatively quick to administer and provide a wealth of additional information, auto-

[6] Since the independent variables were all z-score normalized before using them in the regression model, it is meaningful to compare their weignts directly.

Dependent variable	Mean	Model RMSE	Model RRSE	Model correlation	Features with highest absolute weights (sorted in descending order)
CWPM	129.0	7.1	21.3%	0.977	18,1,17,2
REL_CW	0.93	0.05	70.3%	0.710	1,17,18
MSSA	237.0	11.0	76.6%	0.643	2,3,18,10

Table 6: Dependent variable (DV) means, RMSE, RRSE and correlations for the three linear regression models, as well as lists of features with highest regression weights

Feature number	Feature class	Feature description	Model weight		
			CWPM ($\alpha = 129.0$)	REL_CW ($\alpha = 0.93$)	MSSA ($\alpha = 238.9$)
6	Pronunciation	Normalized AM score	1.12	0.007	-1.17
5	Pronunciation and Accuracy	Normalized ASR confidence score	1.40	0.013	2.41
4	Accuracy	Normalized LM score	-1.32	-0.007	1.24
18	Fluency and Accuracy	Estimated CWPM	20.78	-0.031	6.46
7	Fluency	Rate of long silences	-3.06	-0.017	–
8	Fluency	Mean deviation of long silences	-1.46	-0.013	–
9	Fluency	Rate of repetitions	–	0.009	-1.35
10	Fluency	Total reading time	-3.57	–	-3.06
12	Fluency	Rate of silences (relative to words spoken)	5.04	0.013	–
13	Fluency	Mean of silence duration	2.63	0.010	–
1	Accuracy	Estimated absolute student word errors	19.49	-0.044	–
2	Accuracy	Estimated reading accuracy	10.90	–	-10.52
17	Fluency	Speaking rate	11.92	0.04	–
11	Fluency	Rate of silences (per time unit)	-4.21	–	–
14	Fluency	Mean deviation of silences	-1.50	–	–
3	Accuracy	Estimated REL_CW	–	–	8.27
15	Fluency	Length of uninterrupted phrases (no intervening silences)	–	–	–
16	Fluency	Mean deviation of uninterrupted phrases	–	–	–

Table 7: Frequency of feature usage in three regression models; the feature numbers correspond to the numbers in Table 4

mated scoring of reading proficiency should be considered as an additional assessment for students performing significantly below grade level in reading comprehension. In addition, this measure combined with a measure of audio comprehension of text would provide teachers and administrators with significantly more information about a student's areas of strengths and weaknesses in reading ability beyond what current state assessments can assess.

7. Conclusion and Future Work

In this paper, we have demonstrated that automatically generated reading proficiency scores using features from an automatic speech recognition system, relating to aspects of accuracy, fluency and pronunciation, can yield medium to high correlations (r=0.98 for CWPM, r=0.71 for REL_CW) with equivalent scores based on human error annotations for a diverse middle school population, including students with low reading proficiency and reading disabilities. Furthermore, we confirmed previous findings of significant correlations between reading comprehension scores and oral proficiency scores for our test population, and found correlations between ASR features and reading comprehension scores in a linear regression model of a similar magnitude (r=0.64).

We further demonstrated that while the regression models predicting human scores for oral reading proficiency and reading comprehension include features from all measured areas of oral reading proficiency (fluency, pronunciation, and accuracy), the features with highest regression weights were from the class of reading accuracy and reading speed.

Future work will include exploring the use of additional features in models of oral reading proficiency that can be automatically derived (e.g., prosody), as well as devising methods for automatically determining the point in a text passage where a student's reading sample ended in order to be able to score incompletely read passages with high accuracy, as well.

8. References

[1] U.S. Department of Education, "The nation's report card: Reading," http://nationsreportcard.gov, 2011.

[2] D. Fuchs, S. R. Vaughn, and L. S. Fuchs, *Responsiveness to Intervention*. Newark, DE: International Reading Association, 2008.

[3] L. S. Fuchs, D. Fuchs, and M. K. Hosp, "Oral reading fluency as an indicator of reading competance: A theoretical, empirical, and historical analysis," *Scientific Studies of Reading*, vol. 5, no. 3, pp. 239–256, 2001.

[4] S. L. Deno and D. Marston, "Curriculum-based measurement of oral reading: An indicator of growth in fluency," in *What research has to say about fluency instruction*, S. J. Samuels and A. E. Farstrup, Eds. Newark, DE: International Reading Association, 2006, pp. 179–203.

[5] J. Hasbrouck and G. A. Tindal, "Oral reading fluency norms: A valuable assessment tool for reading teachers," *The Reading Teacher*, vol. 59, no. 7, pp. 636–644, 2006.

[6] M. M. Wayman, T. Wallace, H. I. Wiley, R. Ticha, and C. A. Espin, "Literature synthesis on curriculum-based measurement in reading," *The Journal of Special Education*, vol. 41, no. 2, pp. 85–120, 2007.

[7] R. W. Woodcock, K. S. McGrew, and N. Mather, *Woodcock-Johnson III*. Itasca, IL: Riverside Publishing, 2001.

[8] Wechsler Individual Achievement Test, 2nd Edition, *Woodcock-Johnson III*. London: The Psychological Corp., 2005.

[9] S. L. Deno, "Curriculum-based measurement: The emerging alternative," *Exceptional Children*, vol. 42, pp. 210–232, 1985.

[10] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, pp. 83–93, 2000.

[11] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," *Speech Communication*, vol. 30, pp. 109–119, 2000.

[12] J. Bernstein, A. V. Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.

[13] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 1994.

[14] J. Mostow, J. Beck, S. V. Winter, S. Wang, and B. Tobin, "Predicting oral reading miscues," in *Proceedings of ICSLP*, 2002.

[15] S. Sitaram and J. Mostow, "Mining data from Project LISTEN's Reading Tutor to analyze development of children's oral reading prosody," in *Proceedings of the 25th Florida Artificial Intelligence Research Society*, 2012.

[16] M. Duong, J. Mostow, and S. Sitaram, "Two methods for assessing oral reading prosody," *ACM Transactions on Speech and Language Processing*, vol. 7, no. 4, pp. 11–22, 2002.

[17] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: the role of multiple information sources," in *Proceedings of IEEE international workshop on multimedia signal processing*, 2007.

[18] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation verification of children's speech for automatic literacy assessment," in *Proceedings of InterSpeech*, 2006.

[19] J. Tepperman, M. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, and A. Alwan, "A bayesian network classifier for word-level reading assessment," in *Proceedings of ICSLP*, 2007.

[20] J. Tepperman, M. Gerosa, and S. S. Narayanan, "A generative model for scoring children's reading comprehension," in *Proceedings of the Workshop on Child, Computer, and Interaction*, 2008.

[21] K. Zechner, J. Sabatini, and L. Chen, "Automatic scoring of children's read-aloud text passages and word lists," in *Proceedings of the NAACL-HTL workshop on Innovative Use of NLP for Building Educational Applications*, 2009.

[22] CSLU, "Kids speech corpus," http://www.cslu.ogi.edu/corpora/kids, 2008.

[23] LDC, "The CMU Kids Corpus," http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC97S63, 1997.

[24] I. H. Witten, E. Frank, and M. Hall, *Data mining: Practical machine learning tools and techniques, 3rd Edition*. Burlington, MA: Morgan Kaufmann, 2011.

[25] J. Balogh, J. Bernstein, J. Cheng, A. V. Moere, B. Townshend, and M. Suzuki, "Validation of automated scoring of oral reading," *Educational and Psychological Measurement (online)*, 2011.