Automated Assessment of Oral Reading Prosody

Jack MOSTOW and Minh DUONG
Project LISTEN, Carnegie Mellon University, Pittsburgh, PA, USA[1]

Abstract. We describe an automated method to assess the expressiveness of children's oral reading by measuring how well its prosodic contours correlate in pitch, intensity, pauses, and word reading times with adult narrations of the same sentences. We evaluate the method directly against a common rubric used to assess fluency by hand. We also compare it against manual and automated baselines by its ability to predict fluency and comprehension test scores and gains of 55 children ages 7-10 who used Project LISTEN's Reading Tutor. It outperforms the human-scored rubric, predicts gains, and could help teachers identify which students are making adequate progress.

Keywords. Oral reading, children, fluency, assessment, speech, prosody

1. Introduction

Assessment of children's oral reading is important for multiple reasons – to compare fluency against expected norms [1], provide motivational feedback on rereading to improve fluency [2], analyze the longitudinal development of fluency [3], compare the efficacy of different types of reading practice [4, 5], study the relation of fluency to comprehension [6], and even estimate the reader's comprehension of a given text [7].

Oral reading fluency is the ability to "read text with speed, accuracy, and proper expression" [8, p. 3-1]. Educators measure oral reading fluency in two ways. Oral reading rate is the number of words read correctly per minute. This measure is quick and easy to administer, and correlates strongly with children's comprehension test scores [9]. However, it ignores expressiveness.

Fluency rubrics [e.g., 10] score reading more subjectively and qualitatively against specified criteria. One such rubric [11] rates expression, phrasing, smoothness, and pace on separate 4-point scales.

Previous work on automated assessment of oral reading has focused on oral reading rate [12] or closely related variants such as average inter-word latency [13, 14] or word reading time [5]. Some work [15] has assessed non-native speakers' oral language. That problem is related to oral reading fluency but very different, because it measures the ability to translate thoughts into language, rather than text into speech.

In this paper we address the problem of assessing oral reading expressiveness automatically. Solving this problem would make it possible to assess reading more

[1] The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A0628. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. We also thank Dr. Paula Schwanenflugel for her expertise, and the educators, students, and LISTENers who helped generate our data.

richly and informatively than oral reading rate, and more precisely and consistently than human-scored rubrics. It could more sensitively detect improvement in oral reading, whether across successive rereadings of the same practice text, or in the ability to read unpracticed text fluently. It could serve as the basis for giving children feedback on their oral reading, describing how to read more expressively. It might even enable a tutor to unobtrusively gauge a student's comprehension of a given text without interrupting to test it by asking questions.

The rest of this paper is organized as follows. Section 2 describes our approach. Section 3 evaluates it against various baselines. Section 4 concludes by summarizing contributions and relating them to prior and future work.

2. Approach

Our approach is inspired by previous analyses of children's oral reading prosody, based on the insight that the more expressive a child's reading of a text, the more its prosody tends to resemble fluent adult reading of the same text. Schwanenflugel et al. [6, 16, 17] analyzed adults' and children' readings of the same short text. They painstakingly hand-aligned the text to a spectrogram of each reading to compute the duration of pauses between sentences, at commas after phrases, and in mid-phrase, called "pausal intrusions;" the number of pausal intrusions; the drop in F_0 (pitch) at the ends of sentences; and the intonation contours of the first three sentences, which they computed by measuring F_0 at the vocalic nucleus of each word. Averaging these values across 34 adults yielded a profile of expressive reading. Correlating this profile against the corresponding values for each child quantified the expressiveness of the child's oral reading, and its changes from the end of grade 1 to the end of grade 2, so as to relate them to scores and gains on reading tests administered at those points. F_0 match and the number of pausal intrusions were the best indicators of prosodic change between first and second grades.

Our goal was to analyze the expressiveness of children's assisted oral reading in Project LISTEN's Reading Tutor, which listens to children read aloud, and helps them learn to read [18]. The Reading Tutor and the child take turns choosing what to read from a collection of several hundred stories with recorded adult narrations. The Reading Tutor displays text incrementally, adding a sentence at a time. It uses an automatic speech recognizer (ASR) [19] to listen to the child read the sentence aloud, tracking the child's position in the text to detect deviations from it and identify the start and end points of each word and silence in the recorded oral reading [20]. It responds with spoken and graphical feedback to hesitations and miscues detected by the ASR, as well as the child's requests for help by clicking on hard words.

To automate our analyses we adapted Schwanenflugel et al.'s approach, with several key differences in data, methods, and features. Our data came from the 2005-2006 version of the Reading Tutor:

- We used the Reading Tutor's single adult narration of each sentence, rather than multiple adult readings of it.
- We used children's oral reading recorded over a whole semester of using the Reading Tutor, rather than briefly by researchers for an experiment.
- We administered tests at start and end of semester, rather than a year apart.
- We computed prosodic contours on thousands of sentences, not just a few.

- We used whatever a child happened to read, rather than one common text.

We replaced manual methods with automated ones:

- We used the Reading Tutor's ASR-based time-alignments of text to oral reading, rather than identifying each word's start and end time by hand.
- We used F_0 computed by a pitch tracker [21] and averaged over each entire ASR-aligned word, rather than hand-measured at its manually located vocalic nucleus in a spectrogram.

We used a somewhat different set of features:

- We computed contours for latency, duration, and intensity, not just pitch.
- We computed pitch variability as standard deviation of F_0 rather than as difference between high and low F_0 values, which outliers can distort.
- We computed the latency preceding each word, the production time to say the word, and their sum as the reading time, both absolute and normalized by word length as time per letter, not just the absolute duration of pauses between sentences, at phrase-final commas, and in mid-phrase intrusions.

Thus we represent a prosodic contour of a read sentence as the sequence of values, one for each word in the sentence, of a prosodic feature such as the word's latency, duration, mean pitch, or mean intensity. To quantify the similarity of the child's contour to the adult contour for the same sentence, we simply use its correlation, so we get separate similarity scores for each prosodic feature.

Representing adult and child contours as sequences of word-level values lets us compare them despite differences between their time alignments. Using correlations to score the similarity of adult and child contours serves to factor out differences between adults and children in baseline F_0 (adults' lower pitch), individual variations in intensity, and fluctuations in recording conditions from one sentence to another.

Prosodic features may be undefined for some words. For instance, inter-word latency is undefined for the first word of a sentence, because the Reading Tutor displays one sentence at a time. Prosodic features are undefined for words the Reading Tutor gave help on or did not accept as read correctly.

We adjust for such missing values as follows. If a prosodic feature is undefined for one or more words in an adult or (more typically) child contour, we exclude them, correlate the rest, and penalize the resulting score by the proportion of undefined values.

For each student we average each word- and sentence-level feature over all of that student's reading. In averaging similarity scores over multiple sentences, we weight the average by the number of words on which each score is based. It is not enough simply to weight each sentence by the number of words with defined values, because undefined values in a child's contour indicate mismatches with the adult contour. Thus adult and child contours for a 10-word sentence with half its values missing earn half the similarity score as complete contours for a 5-word sentence, even if they have the same 5 defined values. To avoid outliers, we considered only the 55 students who had at least 20 sentences.

Besides averaging each feature per student, we want to characterize how it evolved over the course of the semester in which the student used the Reading Tutor. We plot each individual observed feature value against the point in time when it was observed, and regress feature values against time. We use the linear regression coefficient for each feature, weighted by its correlation coefficient with time to estimate its rate of change, and use these rates as additional features.

Given a set of features, we predict a dependent variable by using stepwise linear regression to fit statistical models, and selecting the model with maximum adjusted R^2. We use cross-validation to estimate this model's predictive accuracy on unseen data.

3. Evaluation

We evaluated our approach directly against a human-scored fluency rubric, and indirectly by its ability to predict test scores and gains.

3.1. Estimate rubric scores

To analyze the feasibility of automating a conventional multi-dimensional fluency rubric, we took a sample of 10 students in our data set, stratified by oral reading rate. We took the first 10 sentences each from the first and last stories each student read, so as to include variation caused by change in fluency over time. Two independent annotators hand-rated students' expression, phrasing, smoothness, and pace for each sentence on a 4-point scale specified by the fluency rubric [11]. For example, the scale for pace ranges from 1 for "slow and laborious" to 4 for "consistently conversational." We averaged the two judges' ratings over the 20 sentences for each student. To quantify the inter-rater reliability of these two continuous-valued average ratings, Table 1 shows the intraclass correlation coefficient [22] and mean absolute difference for each dimension of the rubric. Agreement was good for every dimension except smoothness, whose scoring was impeded by two factors. First, the Reading Tutor presented text one sentence at a time, which prevented observation of pauses before sentences. Second, the annotators used a tool [23] that played back oral reading one utterance at a time rather than entire sentences, making some aspects harder to score.

It is important to note that inter-rater reliability was considerably lower for scoring *individual sentences* on a discrete 4-point scale, with Kappa only 0.087 for smoothness and ranging from 0.225 to 0.297 for the other dimensions. Possibly, reading professionals would agree better, or the rubric is simply too subjective. However, the simplest explanation is that a single sentence is insufficient to rate reliably.

Table 1: Inter-rater agreement for the fluency rubric at the student level

Rubric dimension	**Consistency (ICC)**	**Mean absolute difference**
Expression and Volume	0.798	0.287
Phrasing	0.681	0.408
Smoothness	0.283	0.674
Pace	0.721	0.387

We predicted the average of the two annotators' ratings from static features of each student's oral reading (excluding rate-of-change features), using SPSS stepwise linear regression on the rest of the students. As Table 2 shows, predicted values correlated significantly ($p<.05$) with these ratings for expression and smoothness.

Table 2: Predicting rubric scores using automated features

Rubric dimension	**Cross-validated correlation**
Expression and Volume	0.902
Phrasing	0.281
Smoothness	0.768
Pace	0.448

3.2. Predict test scores

We also wanted to test the feasibility of predicting students' scores and gains on standard measures of fluency and comprehension. We measured gains as posttest (end-of-semester) minus pretest (before using the Reading Tutor). The fluency test measured oral reading rate as the number of words read correctly in one minute on a passage at the student's grade level (2, 3, or 4) [9]. Word reading time correlates strongly with word length, and the test passages for the different grades differed in average word length (3.8, 4.1, and 4.4, respectively). To control for this difference, we normalized fluency test scores as letters per second instead of words per minute. The comprehension test was the Passage Comprehension component of the Woodcock Reading Mastery Test, consisting of short passages with multiple choice questions.

Beck et al. [14] introduced a model using features based on inter-word latency. This *latency-based* model predicted posttest fluency well, with cross-validated mean within-grade correlation of 0.83. We extended it into a *correlation+latency* model by adding our correlation-based features of oral reading, including rate-of-change features as well as static features. We trained models on all the data from the 55 students.

In general, pretest scores are strong predictors of posttest scores, so we compared correlation+latency models against pretest scores as predictors of students' posttest scores and pre- to posttest gains. We also trained hybrid *correlation+latency+pretest* models to predict posttest scores and gains by augmenting pretest scores with the automated features, so as to measure the additional predictive power the features added.

We evaluated models by their cross-validated mean within-grade correlations between predicted and actual values. We correlated predicted with actual values for each grade, and computed the arithmetic mean of these within-grade correlations. As Table 3 shows, the correlation+latency models outperformed pretest scores and latency-based models across the board in predicting fluency and comprehension scores and gains – in two cases even better than after adding pretest scores.

Table 3: Cross-validated mean within-grade correlations between predicted and actual values

Model	Posttest fluency	Comp. posttest	Fluency gain	Comp. gain
Pretest scores	0.809	0.738	-0.741	0.202
Latency-based	0.859	0.724	0.524	0.453
Correlation+latency	0.872	0.763	0.934	0.504
Correl.+latency+pretest	0.965	0.690	0.867	0.643

In our correlation+latency model, the features that explained the most variance in posttest fluency were, in order, percentage of words read without hesitation, normalized production time, word reading time, normalized production change rate, and latency correlation with adult narrations. Latency strongly influences 3 of our 5 top features (including word reading time, which is the sum of latency and production). The top predictors in our correlation+latency model of posttest comprehension scores were the percentage of words with defined latency (i.e., read correctly and without omitting the previous word [14]), latency correlation, pitch correlation, percentage of Dolch (high-frequency) words with minimal latency, and change rate of normalized production correlation.

In comparison, Miller and Schwanenflugel [17] found pitch correlation to be the best prosodic predictor of grade 3 fluency, and reduction in pausal intrusions from grade 1 to grade 2 to be the best at predicting grade 3 comprehension. Latency and pitch correlation were likewise among our most predictive features – but latency predicted fluency, and pitch correlation predicted comprehension, rather than vice versa. Disfluency hinders comprehension, but one expects intonation to reflect understanding.

3.3. Compare to rubric-based prediction

How did our correlation-based features compare to the fluency rubric in predicting the same students' test scores? To find out, we used the 10 students with rubric scores as a held-out set. We fit correlation-based, pretest-based, and correlation+pretest models to the 45 remaining students' posttest scores, and tested on the held-out data for the 10-student sample. We compared them to models trained on the hand-scored rubric ratings for the 10 students using Weka's linear regression with the M5 feature selection method, and evaluated models using leave-1-out cross-validation to avoid overfitting.

Table 4 compares correlations for the four types of model. All models predicted posttest fluency with correlation 0.93 or better, thanks to the heterogeneity of a small sample of students stratified by fluency. In predicting posttest comprehension, the correlation-based model performed better than the rubric-based model. But was it because its features were actually more predictive, or just that it used more data?

Table 4: Correlation between predicted and actual posttest scores of 10 rubric-scored students

Model	Posttest fluency	Posttest comprehension
Pretest-based	0.95	0.79
Correlation-based	0.95	0.45
Correlation+pretest	0.97	0.66
Rubric-based	0.93	0.30

3.4. Vary the amount of training data

To distinguish whether the inferior performance of the rubric-based models was due to worse features or to less data, we compared them against correlation-based models trained on the same 10 students. One training condition used all their data. Another condition used only the 200 sentences scored by the rubric annotator. Table 5 compares the leave-out-one cross-validated correlation of both models to the feature-based models described earlier. As expected, using less data from the same 10 students resulted in less competitive models, for both fluency and comprehension. The models trained on all sentences from the 10 students also performed better than the models trained on the rest of the students, which suggests that they still benefitted from information about the test set, even though they were cross-validated.

Table 5: Predictive accuracy of correlation-based models trained on different amounts of data

Training data for correlation-based model	Posttest fluency	Posttest comprehension
Trained on all 53 other students	0.95	0.45
Trained on all sentences from 10 students	0.98	0.71
Trained on only 200 sentences from 10 students	0.90	0.34

4. Contributions, Relation to Prior Work, and Future Work

In this paper we have presented a method to assess oral reading prosody automatically, and evaluated it on a large corpus of real data recorded by Project LISTEN's Reading Tutor in real schools by real students. We evaluated both directly against a human-scored rubric, and indirectly by predicting scores and gains on standard tests of oral reading fluency and comprehension.

The resulting estimates compared favorably to alternatives. Although the method did poorly at estimating fluency rubric scores for individual sentences, where inter-rater reliability was low, it did better at estimating scores at the student level, and out-performed rubric scores in predicting students' fluency posttest scores. It predicted gains in fluency and comprehension much better than standard pretest scores did.

Our work differs from previous automated assessments of oral reading fluency in what and how they estimate. Balogh et al. [12] used a proprietary system to score the speed and accuracy of adults' oral reading. They validated against human judges counting the number of words read correctly in the same recorded readings. The human judges correlated strongly (0.96-1.00) with the automated assessment – as strongly as they did with each other. It is important to note that individuals' reading rates fluctuate for many reasons, so their rates on different passages or even the same passage at different times correlate less than perfectly with each other. Thus predicting posttest fluency is inherently harder than measuring reading rate on the same recording.

Beck et al. [14] used various aggregate features of word latencies and help requests by 87 children in grades 1-4 over a two-month window in Project LISTEN's Reading Tutor to predict their fluency test scores. They did not predict comprehension or gains, but we did so by replicating their model, and improved on its predictive accuracy by incorporating additional features inspired by the work of Schwanenflugel et al.

Zhang et al. [7] attempted to detect moment-to-moment fluctuations in children's comprehension of the text they were reading aloud in the Reading Tutor. They trained a model to predict performance on multiple-choice comprehension questions inserted during the reading. Oral reading behavior improved model fit only marginally after controlling for student identity and question attributes affecting difficulty. Only oral reading features related to oral reading rate and accuracy achieved significance – but none of their prosodic features compared prosodic contours to fluent narrations.

This paper advances over previous analyses of oral reading prosody [6, 16, 17] in multiple directions. We replaced manual with automated analysis. We scaled up from three sentences to thousands of sentences. We extended the set of features to include intensity and word reading time as well as pitch and pauses. We eliminated the requirement for many adults to read each sentence in order to average their prosodic contours; we need only one adult narration per sentence. We eliminated the requirement for all children to read the same text in order to compare them; we evaluate prosody on whatever a child read. We eliminated the requirement to take time out for testing, by instead using data already captured during routine use of the Reading Tutor. Not only is there more such data per student, it is longitudinal, allowing us to measure features' change over time, and predict scores and gains better. Raising ASR accuracy, adding features of the text (e.g. syntax and punctuation) and student (e.g. previous encounters of each word), or refining aspects of the data analysis (e.g. outlier filtering and model form) may further enhance the valuable ability to monitor children's reading growth continually as a byproduct of improving it in a tutor.

References (Project LISTEN publications are at www.cs.cmu.edu/~listen)

[1] Hasbrouck, J.E. and G.A. Tindal. Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. *The Reading Teacher*, 2006. *59*(7): p. 636-644.

[2] Kuhn, M.R. and S.A. Stahl. Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 2003. *95*(1): p. 3–21.

[3] O'Connor, R.E., A. White, and H.L. Swanson. Repeated reading versus continuous reading: Influences on reading fluency and comprehension. *Exceptional Children*, 2007. *74*(1): p. 31-46.

[4] Kuhn, M.R., P.J. Schwanenflugel, R.D. Morris, L.M. Morrow, D.G. Woo, E.B. Meisinger, R.A. Sevcik, B.A. Bradley, and S.A. Stahl. Teaching Children to Become Fluent and Automatic Readers. *Journal of Literacy Research*, 2006. *38*(4): p. 357-387.

[5] Beck, J.E. and J. Mostow. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. *9th International Conference on Intelligent Tutoring Systems*, 353-362. Nominated for Best Paper. 2008. Montreal.

[6] Schwanenflugel, P.J., E.B. Meisinger, J.M. Wisenbaker, M.R. Kuhn, G.P. Strauss, and R.D. Morris. Becoming a fluent and automatic reader in the early elementary school years. *Reading Research Quarterly*, 2006. *41*(4): p. 496-522.

[7] Zhang, X., J. Mostow, and J.E. Beck. Can a Computer Listen for Fluctuations in Reading Comprehension? *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, 495-502. 2007. Marina del Rey, CA: IOS Press.

[8] NRP. Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. 2000, National Institute of Child Health & Human Development. At http://www.nichd.nih.gov/publications/nrppubskey.cfm: Washington, DC.

[9] Deno, S.L. Curriculum-Based Measurement: The emerging alternative. *Exceptional Children*, 1985. *52*(3): p. 219-232.

[10] Pinnell, G.S., J.J. Pikulski, K.K. Wixson, J.R. Campbell, P.B. Gough, and A.S. Beatty. Listening to Children Read Aloud: Oral Reading Fluency. 1995, National Center for Educational Statistics: Washington, DC.

[11] Zutell, J. and T.V. Rasinski. Training Teachers to Attend to Their Students' Oral Reading Fluency. *Theory into Practice*, 1991. *30*(3): p. 211-17.

[12] Balogh, J., J. Bernstein, J. Cheng, and B. Townshend. Automatic Evaluation of Reading Accuracy: Assessing Machine Scores. *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE)* 2007. Farmington, PA.

[13] Mostow, J. and G. Aist. The sounds of silence: Towards automated evaluation of student learning in a Reading Tutor that listens. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, 355-361. 1997. Providence, RI: American Association for Artificial Intelligence.

[14] Beck, J.E., P. Jia, and J. Mostow. Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning*, 2004. *2*(1-2): p. 61-81.

[15] Bernstein, J., M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub. Automatic evaluation and training in English pronunciation. *International Conference on Speech and Language Processing (ICSLP-90)* 1990. Kobe, Japan.

[16] Schwanenflugel, P.J., A.M. Hamilton, M.R. Kuhn, J.M. Wisenbaker, and S.A. Stahl. Becoming a Fluent Reader: Reading Skill and Prosodic Features in the Oral Reading of Young Readers. *Journal of Educational Psychology*, 2004. *96*(1): p. 119-129.

[17] Miller, J. and P.J. Schwanenflugel. A Longitudinal Study of the Development of Reading Prosody as a Dimension of Oral Reading Fluency in Early Elementary School Children. *Reading Research Quarterly*, 2008. *43*(4): p. 336–354.

[18] Mostow, J., G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, M.B. Sklar, and B. Tobin. Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 2003. *29*(1): p. 61-117.

[19] CMU. The CMU Sphinx Group Open Source Speech Recognition Engines [software at http://cmusphinx.sourceforge.net]. 2008.

[20] Mostow, J., S.F. Roth, A.G. Hauptmann, and M. Kane. A prototype reading coach that listens [AAAI-94 Outstanding Paper]. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 785-792. 1994. Seattle, WA: American Association for Artificial Intelligence.

[21] Boersma, P. and D. Weenink. Praat: doing phonetics by computer (Version 5.0.33) [Computer program downloaded from http://www.praat.org/]. 2008.

[22] Shrout, P.E. and J.L. Fleiss. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 1979. *86*(2): p. 420-428.

[23] Mostow, J., J. Beck, and others. Lessons from Project LISTEN's Session Browser. In C.R. Morales, et al., Editors, *Handbook of Educational Data Mining*. Taylor & Francis Group: London, under review.