Two Methods for Assessing Oral Reading Prosody

MINH DUONG, Carnegie Mellon University
JACK MOSTOW, Carnegie Mellon University
SUNAYANA SITARAM, Carnegie Mellon University

We compare two types of models to assess the prosody of children's oral reading. Template models measure how well the child's prosodic contour in reading a given sentence correlates in pitch, intensity, pauses, or word reading times with an adult narration of the same sentence. We evaluate template models directly against a common rubric used to assess fluency by hand, and indirectly by their ability to predict fluency and comprehension test scores and gains of 10 children who used Project LISTEN's Reading Tutor; the template models outpredict the human assessment.

We also use the same set of adult narrations to train generalized models for mapping text to prosody, and use them to evaluate children's prosody. Using only durational features for both types of models, the generalized models perform better at predicting fluency and comprehension posttest scores of 55 children ages 7-10, with adjusted R2 of 0.6. Such models could help teachers identify which students are making adequate progress. The generalized models have the additional advantage of not requiring an adult narration of every sentence.

Categories and Subject Descriptors: **I 2.7 [Artificial Intelligence**]: Natural language processing – Speech recognition and synthesis; **K.3.1 [Computers and Education]:** Computer Uses in Education – Computer-assisted instruction (CAI).

General Terms: Algorithms, Experimentation, Human Factors, Measurement

Additional Key Words and Phrases: Oral reading fluency, prosody, assessment, intelligent tutoring system, children

ACM Reference Format:

DUONG, M., MOSTOW, J. and SITARAM, S. 2011. Two methods for assessing oral reading prosody *ACM Transactions on Speech and Language Processing* (Special Issue on Speech and Language Processing of Children's Speech for Child-machine Interaction Applications).

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. The first author participated in this research as a graduate student at Carnegie Mellon Language Technologies Institute.

Authors' addresses: Minh Duong, Facebook, 1601 S. California Ave., Palo Alto, CA 94304, USA, E-mail: mnduong@cs.cmu.edu; Jack Mostow, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA, E-mail: mostow@cs.cmu.edu; Sunayana Sitaram, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 15213 USA, E-mail: ssitaram@cs.cmu.edu.

1. INTRODUCTION

Assessment of children's oral reading is important for multiple reasons – to compare fluency against expected norms [Hasbrouck and Tindal, 2006], provide motivational feedback on rereading to improve fluency [Kuhn and Stahl, 2003], analyze the longitudinal development of fluency [O'Connor et al., 2007], compare the efficacy of different types of reading practice [Beck and Mostow, 2008; Kuhn et al., 2006], study the relation of fluency to comprehension [Schwanenflugel et al., 2006], and even estimate the reader's fluctuating comprehension of a given text [Zhang et al., 2007].

Table 1: Fluency rubric adapted from [Zutell and Rasinski, 1991]; downloaded from www.timrasinski.com/presentations/multidimensional_fluency_rubric_4_factors.pdf

Dimension	1	2	3	4
Expression and Volume	Reads as if just trying to "get words out." Little sense of trying to make text sound like natural language. Tends to read in a quiet voice.	Begins to use voice to make text sound like natural language in some areas but not in others. Focus remains largely on pronouncing words. Still reads in a quiet voice.	Makes text sound like natural language throughout the better part of the passage. Occasionally slips into expressionless reading. Voice volume is generally appropriate throughout the text.	Reads with good expression and enthusiasm throughout the text. Varies expression and volume to match his or her interpretation of the passage.
Phrasing	Reads in monotone with little sense of phrase boundaries; frequently reads word-by-word	Frequently reads in two and three word phrases, giving the impression of choppy reading; improper stress and intonation fail to mark ends of sentences and clauses	Reads with a mixture of run-ons, mid sentence pauses for breath, and some choppiness; reasonable stress and intonation.	Generally reads with good phrasing, mostly in clause and sentence units, with adequate attention to expression.
Smoothness	Makes frequent extended pauses, hesitations, false starts, sound-outs, repetitions, and/or multiple attempts.	Experiences several "rough spots" in text where extended pauses or hesitations are more frequent and disruptive.	Occasionally breaks smooth rhythm because of difficulties with specific words and/or structures.	Generally reads smoothly with some breaks, but resolves word and structure difficulties quickly, usually through self-correction.
Pace	Reads slowly and laboriously.	Reads moderately slowly.	Reads with an uneven mixture of fast and slow pace.	Consistently reads at a conversational pace; appropriate rate throughout reading.

Oral reading fluency is the ability to "read text with speed, accuracy, and proper expression" [NRP, 2000]. Expressiveness is the ability to read with a prosodic contour appropriate to the meaning of the text, which reflects an understanding of the text [Schwanenflugel et al., 2006]. 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 [Deno, 1985]. However, oral reading rate ignores expressiveness.

Fluency rubrics [e.g., Pinnell et al., 1995] score reading more subjectively and qualitatively against specified criteria. One widely used fluency rubric, shown in Table 1, rates expression, phrasing, smoothness, and pace on separate 4-point scales. For example, the scale for pace ranges from 1 for slow and laborious to 4 for consistently conversational.

In this paper we address the problem of assessing oral reading prosody automatically. Solving this problem would make it possible to assess reading more 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 re-readings 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, telling them how to read more expressively. It might even enable a tutor to gauge a student's comprehension of a given text unobtrusively [Zhang et al., 2007], without interrupting to insert multiple-choice probes [Mostow et al., 2004] or ask open-ended questions [Gerosa and Narayanan, 2008].

Our data consist of children's oral reading assisted and recorded by Project LISTEN's Reading Tutor, which listens to children read aloud, and helps them learn to read [Mostow et al., 2003]. 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 the Sphinx automatic speech recognizer (ASR) [CMU, 2008] 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 [Mostow et al., 1994]. It responds with spoken and graphical feedback when the ASR detects hesitations or miscues, or when the child clicks for help on hard words or sentences. The spoken feedback uses recorded human speech, including a fluent adult narration of each sentence, time-aligned against the sentence text by forced alignment.

We present two types of models for assessing children's oral reading prosody. Both models score oral reading one sentence at a time. A *template model* scores the child's reading of a sentence by its similarity to the adult narration of the same sentence, based on whatever prosodic contour the sentence led the narrator to produce. Thus it may implicitly capture subtle text features to which a human narrator is sensitive. However, it cannot evaluate unnarrated sentences. A *generalized model* is trained on a corpus of adult narrations, and scores the child's reading against the trained model instead of assuming that an adult narration is available for that sentence.

Moreover, generalized models are trained on multiple adult voices, so they are not specific to the idiosyncrasies of any one speaker. To determine whether such generalization is possible without sacrificing the accuracy of the sentence-specific approach, we compare the models' ability to predict students' fluency and comprehension test scores.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes template models and Section 4 evaluates them against various baselines. Sections 5 and 6 respectively describe and evaluate generalized models. Section 7 concludes by summarizing contributions and proposing future work.

2. RELATION TO PRIOR WORK

Previous work has assessed children's oral reading at different levels of granularity. Some work, especially to assess pronunciation by native and non-native speakers, has focused on reading individual words, either in isolation [Tepperman et al., 2011] or in lists [Price et al., 2009]. Other work has focused on reading connected text, using speech recognition in automated tutors to track the reader's position and detect miscues [Beattie, 2010; Hagen et al., 2007; Mostow et al., 1994]. Assessment of connected reading has focused on accuracy of word identification [Banerjee et al., 2003; Black et al., 2007; Lee et al., 2004], mastery of grapheme-to-phoneme mappings [Beck and Sison, 2006], and fluency of oral reading – more specifically, oral reading rate [Balogh et al., 2007] or closely related variants such as average inter-word latency [Beck et al., 2004; Mostow and Aist, 1997] or word reading time [Beck and Mostow, 2008]. Some work [Bernstein et al., 1990; Tepperman et al., 2007] has assessed (adult) 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.

Our work differs from previous automated assessments of oral reading fluency in what and how they estimate. Balogh et al. [2007] 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. [2004] 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. However, they did not predict comprehension scores or gains.

Zhang et al. [2007] 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.

3. TEMPLATE MODELS

We now describe how we represent and compare prosodic contours, and the manual analysis methodology that inspired our approach.

3.1 Representing prosodic contours

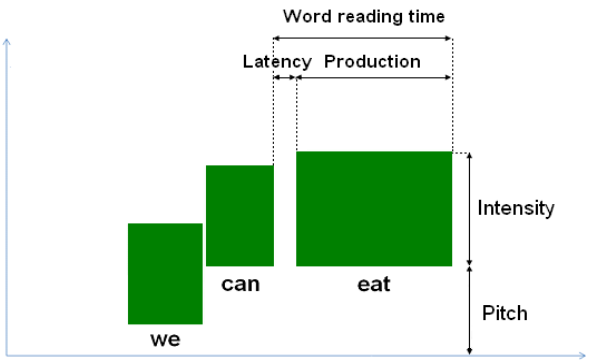
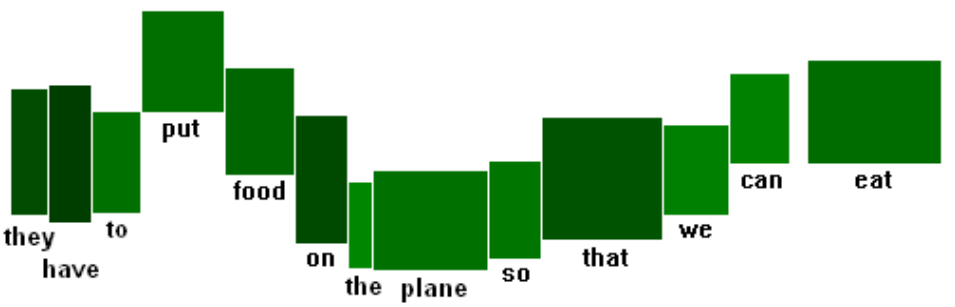
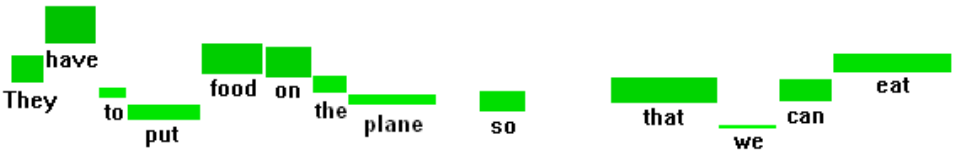


Figure 1: Visualization of a prosodic contour

"Expressiveness" is important but defined imprecisely and judged subjectively. Therefore we use a more precisely defined construct: prosodic contour. 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. As Figure 1 shows, we can visualize a prosodic contour as a sequence of rectangles, displaying prosodic features of each read word as graphical features of the rectangle representing it: duration as width, pitch as vertical position, and intensity as height and color. Figure 2a, b, and c visualize prosodic contours for readings of the same sentence by an adult narrator, a fluent child, and a disfluent child. The contour for the disfluent reading reflects its longer duration and frequent pauses.



a. Fluent adult narration



b. Fluent child reading

c. Disfluent child reading

Figure 2: 3 prosodic contours for "They have to put food on the table so that we can eat"

3.2 Scaling up manual analysis of oral reading prosody

The template model 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 and her collaborators [Miller and Schwanenflugel, 2006; Miller and Schwanenflugel, 2008; Schwanenflugel et al., 2004; Schwanenflugel et al., 2006] 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 prosodic profile of 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 (correlation to adult intonational contour) and the number of pausal intrusions were the best indicators of prosodic change between first and second grades.

Our goal was to analyze the prosody of children's assisted oral reading in Project LISTEN's Reading Tutor. Our data came from the 2005-2006 version of the Reading Tutor. To automate our analysis we adapted Schwanenflugel et al.'s approach, with several key differences in data, methods, and features:

- We used the Reading Tutor's single adult narration of each sentence, rather than multiple adult readings of it. Although a sentence can be read fluently in more than one way, Schwanenflugel (personal email communication, 10/18/2008) found "extremely high correlation among adults in our sample, even across regions."
- 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 a single 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 estimated F_0 using Praat [Boersma and Weenink, 2008] and averaged F_0 over the time interval occupied by each word, rather than hand-measuring F_0 at the word's manually located vocalic nucleus in a spectrogram.
- We computed pitch variability as standard deviation of F_0 rather than as difference between high and low F_0 values, because standard deviation is more robust to outliers than range is.

We used a somewhat different set of word-level features, listed in Table 2:

- We computed contours for word latency, duration, and intensity, not just pitch.
- We computed production time to say each word, not just the latency preceding each word, and total reading time for each word as latency plus production time.
- We computed latency, production, and reading time not only as absolute durations, but also normalized by word length as time per letter.
- We computed the latency preceding every word of each sentence except the first word, not just at phrase-final commas and in mid-phrase intrusions.
- We did not compute the latency of pauses between sentences, because the Reading Tutor displayed text incrementally, adding one sentence at a time; consequently inter-sentential pauses included time waiting for the Reading Tutor to give feedback (if any) and then display the next sentence.

Table 2: Word-level features used by template model

Raw reading time	Normalized reading time
Raw latency	Normalized latency
Raw production	Normalized production

3.3 Comparing prosodic contours

To quantify the similarity of the child's oral reading to an adult narration of the same sentence, we adopted Schwanenflugel et al.'s metric: simply compute the correlation between the two contours. We get separate

similarity scores for each prosodic feature. For example, the intensity contours in Figure 2a (adult) and b (fluent child) correlate at 0.77. In contrast, the intensity contours in Figures 2a (adult) and c (disfluent child) correlate at only 0.02.

Representing child and adult contours as sequences of word-level values lets us compare them despite differences between their time alignments. Moreover, using correlations to score the similarity of child and adult contours is an elegant way to normalize them: it serves to factor out differences between children and adults in baseline F_0 (adults' lower pitch), individual variations in intensity, and fluctuations in recording conditions from one sentence to another. An alternative way to normalize pitch and intensity values would use their deviations from baseline values computed by averaging over the utterance or recent utterances.

Table 3 lists three categories of sentence-level features used in template models. Correlating against an adult contour is just one way to score a child's oral reading of a sentence. Many prosodic features are also informative to average over the utterance. Finally, some features are defined at the sentence level but not for individual words.

Table 3: Sentence-level features used by template models

1. Correlation of pitch contour to adult narration
2. Correlation of intensity contour to adult narration
3. Correlation of raw reading time to adult narration
4. Correlation of raw production to adult narration
5. Correlation of raw latency to adult narration
6. Correlation of normalized reading time to adult narration
7. Correlation of normalized production to adult narration
8. Correlation of normalized latency to adult narration
9. Mean raw reading time
10. Mean normalized reading time
11. Mean raw production
12. Mean normalized production
13. Mean raw latency
14. Mean normalized latency
15. Pause frequency
16. Pitch variation

To assess a child's oral reading prosody overall, we average each sentence-level feature over all the sentences read by that child. The read sentences in our corpus averaged 6.7 words long but varied in length, so they should not count equally. In averaging sentence-level features over multiple sentences, we therefore weight the average by the number of words on which each feature is based.

Prosodic features can 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, such as the words *on*, *so*, and *we* in Figure 2c.

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, compute the sentence-level feature by correlating or

averaging over the rest of the words, and penalize the resulting score by the proportion of undefined values.

It is not enough just 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 a child contour with only 5 defined word-level values for a 10-word sentence gets only half of the sentence-level feature value that a complete contour for a 5-word sentence with the same 5 defined word-level values would get.

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, because we care not only about children's reading proficiency, but about their rate of progress over time. We plot the feature values observed for each student against the calendar times when they were observed. We use their linear regression coefficient to estimate the feature's rate of change for that student, and their correlation to measure the strength of their relationship for that student. We multiply the linear regression coefficient by the correlation to obtain the weighted rate-of-change features listed in Table 4.

Table 4: Weighted rate-of-change features:

1. Growth in correlation of total
2. Growth in correlation of pitch
3. Growth in correlation of intensity
4. Growth in correlation of raw reading time
5. Growth in correlation of raw production
6. Growth in correlation of raw latency
7. Growth in correlation of normalized reading time
8. Growth in correlation of normalized production
9. Growth in correlation of normalized latency
10. Growth in mean raw reading time
11. Growth in mean raw production
12. Growth in mean raw latency
13. Growth in mean normalized reading time
14. Growth in mean normalized production
15. Growth in mean normalized latency
16. Growth in pause frequency
17. Growth in help rate
18. Growth in percentage words correct

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.

4. EVALUATION OF TEMPLATE MODELS

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

4.1 Estimating 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 to ensure a diverse sample. 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 shown in Table 1 [Zutell and Rasinski, 1991].

We averaged the two judges' ratings over the 20 sentences for each student on each dimension and overall (averaged across the 4 dimensions). (One judge marked 3 sentences "unscorable" because they were off-task, so we changed their ratings to 1 on each dimension for purposes of analysis.) To quantify the inter-rater reliability of these two continuous-valued average ratings for each dimension of the rubric, Table 5 uses two types of measures. The intraclass correlation coefficient [Shrout and Fleiss, 1979] measures inter-rater reliability on a scale from 0 (worst) to 1 (perfect). Absolute difference on a 4-point scale is less sophisticated statistically but more understandable intuitively, both as an average value and in terms of how often it falls under some threshold.

Intraclass correlation would be 1 for perfect agreement and was approximately 0.7 or better at the student level for every dimension except smoothness, whose scoring was impeded by two factors. First, the Reading Tutor presented text one sentence at a time, which precluded observation of normal pauses before sentences. Second, the annotators used a tool [Mostow et al., 2010] that played back oral reading one utterance at a time rather than an entire sentence, making some aspects harder to score. An utterance is a segment of speech delimited by a silence, so a sentence reading that contains pauses can consist of multiple utterances.

To compare to published results, we also analyzed inter-rater differences as reported by Rasinski et al. [2009] for a variant scale that collapsed Phrasing and Expression: "Exact matches or agreement between coders was achieved in 81% of the samples, and adjacent (one point difference) agreement was found in the remaining 19%." In contrast to their discrete 4-point student-level scores, our student-level scores are averages of sentence-level scores, which seldom match exactly. Therefore, the last two columns of Table 5 report the equivalent reliabilities – the percentages of students for whom these averages differed by less than 0.5 or up to 1. We had lower percentages of agreement within 0.5 points than the 81% rate of exact matches reported by Rasinski et al., but like them we achieved 100% agreement within 1 point on all scales – except smoothness, where the two judges nevertheless differed by less than 1.5 points on all 10 students.

However, even the reliability reported by the rubric author is disappointing. Agreement within one point on a 4-point scale is not a very stringent criterion, considering that two randomly chosen values will achieve it 75% of the time. Moreover, this agreement is at the student level, based on reading an entire passage.

It is therefore unsurprising that inter-rater reliability was considerably lower for scoring individual sentences on a discrete 4-point scale, as

Table 6 shows in comparison to Table 5. Average inter-rater differences were larger for single sentences than averaged over the 20 sentences for each student. Perhaps 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 on this rubric.

Table 5: Inter-rater agreement on fluency rubric at the student level (N=10 students)

Rubric dimension	Inter-rater reliability (Intraclass correlation)	Mean absolute difference	Under 0.5	Within 1
Expression and Volume	0.798	0.290	70%	100%
Phrasing	0.681	0.415	60%	100%
Smoothness	0.283	0.695	30%	80%
Pace	0.721	0.400	50%	100%
Overall	0.685	0.408	60%	100%

Table 6: Inter-rater agreement on fluency rubric at the sentence level (N=200 sentences)

Rubric dimension	Inter-rater reliability (Intraclass correlation)	Mean absolute difference	Under 0.5	Within 1
Expression and Volume	0.593	0.570	50%	94%
Phrasing	0.535	0.685	43%	89%
Smoothness	0.398	0.845	34%	84%
Pace	0.578	0.620	46%	93%
Overall	0.638	0.570	41%	87%

Table 7: Experiments to estimate rubric scores

Model	Fluency rubric
Input	Prosodic features of 200 sentences read by 10 students
Train	Stepwise linear regression
Aggregate	Averages over 20 sentences for each student
Predict	Sentence level rubric scores on four dimensions
Evaluate	Leave-1-out cross-validated correlation

To compute cross-validated correlations, we predicted the average of the two annotators' ratings from static (i.e. excluding rate-of-change) features of each student's oral reading, using SPSS stepwise linear regression on the rest of the students.

Table 7 summarizes this experiment. As Table 8 shows, predicted values correlated significantly ($p<0.01$) [Soper, 2010] with these ratings for expression and smoothness, but not for phrasing and pace. This performance is even less impressive because the sample of 10 students was stratified, and the resulting heterogeneity tends to magnify correlations. In short, direct comparison fared poorly for want of a reliable human gold standard; we

needed a better way to evaluate our approach, so we resorted to an indirect comparison.

Table 8: Estimating student-level rubric scores using automated features (N=10 students)

Rubric dimension	Cross-validated correlation	P value
Expression and Volume	0.902	0.000
Phrasing	0.281	0.431
Smoothness	0.768	0.009
Pace	0.448	0.194

4.2 Predicting test scores

Ideally we would evaluate methods for automated assessment of oral reading prosody directly against a gold standard measure assessing the prosody of each read sentence. The obvious candidate for such a measure is a human-scored rubric to evaluate oral reading fluency, as described in Section 4.1. However, these labor-intensive ratings would cost too much to score large amounts of data. Moreover, their inter-rater reliability on a sample of 200 read sentences was disappointing, especially at the level of individual sentences.

In the absence of a reliable gold standard, we evaluated our automated methods indirectly by how well they predicted students' scores and gains on standard measures of oral reading fluency and comprehension outside the Reading Tutor. We administered highly reliable, psychometrically validated paper tests individually at the beginning (pretest) and end (posttest) of the semester. 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) [Deno, 1985]. To measure oral reading rate, we used grade-appropriate fluency passages provided by reading expert and former LISTENer Dr. Rollanda O'Connor. In prior studies with these test passages the within-grade parallel form reliability was 0.92, consistent with other reliability estimates for oral reading rate [e.g., Compton and Appleton, 2004].

Long words take longer to read: mean word reading time for k-letter words correlates very strongly with k (R^2 = .983) [Mostow and Beck, 2005]. Harder text has longer words, so our test passages for grades 2,3, and 4 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 oral reading rate measured by such fluency tests is not the same construct as the oral reading prosody assessed directly (though unreliably) by human raters. However, previous work [Benjamin and Schwanenflugel, 2010] had already related children's oral reading prosody on two text passages to their comprehension scores. Moreover, ideal assessment of oral reading prosody would unobtrusively measure comprehension of the text.

Thus reading comprehension scores are a compelling criterion to validate against. An even more compelling criterion would be comprehension of the specific sentences read aloud [Zhang et al., 2007]. However, interrupting after each sentence to test comprehension would slow down reading, disrupt its flow, and affect comprehension.

To measure reading comprehension, we used the Passage Comprehension component of the commercially available Woodcock Reading Mastery Test [Woodcock, 1998]. The Passage Comprehension subtest consists of short passages with multiple choice cloze (fill-in-the-blank) questions, and has within-grade reliability of 0.9 in early grades [Berninger et al., 2006].

To evaluate our prosody measures, we needed models to predict test scores from them. Beck et al. [2004] 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, compared to the upper bound of 1. (The overall correlation for a sample that spans multiple grades reflects the heterogeneity of the sample as much as the accuracy of the model. Computing the correlation within each grade and averaging these within-grade correlations avoids this spurious inflation.) Beck et al. did not predict comprehension or gains, but we did so by replicating their model.

We then extended this latency-based model into a *correlation+latency* model by adding our template-based features of oral reading, including rate-of-change features as well as static features. To avoid outliers, we considered only the 55 students who read at least 20 sentences. We trained models on all the data from these 55 students, using stepwise linear regression in SPSS.

Test scores should increase over time. Correlating predicted against actual posttest scores and gains gives credit only for predicting differences among students, not merely for predicting the average increase over pretest scores.

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 contributed by the features.

Table 9 summarizes this experiment. We evaluated models by their cross-validated mean within-grade correlations between predicted and actual values. Like Beck et al. [2004], we correlated predicted with actual values for each grade, and computed the arithmetic mean of these within-grade correlations. As Table 10 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. The first row of the table shows how well each pretest predicted posttest and gains in the same skill. We also found that *fluency* pretest predicted posttest scores and gains in *comprehension* with respective mean within-grade correlations of 0.617 (unsurprisingly, lower than using pretest to predict posttest scores for the same skill) and 0.297 (actually better than comprehension pretest predicted

comprehension gains). We computed these correlations within three grades (2, 3 and 4) with N = 33, 14, and 8 students, respectively. Correlations in grades 3 and 4 were generally higher than in grade 2 and the means reported in Table 10.

Table 9: Experiment to predict test scores and gains

Model	Template model
Input	77,600 sentence readings by 55 students aged 7-10 with test scores; 24,816 sentence narrations by 20 adults
Train	Stepwise linear regression
Aggregate	Average features across sentences, weighted by number of words read
Predict	Posttest fluency and comprehension scores and gains
Evaluate	Leave-one-out cross-validated correlation of actual and predicted values

Table 10: Cross-validated mean within-grade correlations of actual values (N=55 students) to values predicted using regression against different features

Model features	Posttest fluency	Comp. posttest	Fluency gain	Comp. gain
Pretest score	0.809	0.738	-0.741	0.202
Latency	0.859	0.724	0.524	0.453
Correlations+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 accepted as 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., accepted by the speech recognizer as read correctly without omitting the previous word [Beck et al., 2004]), latency correlation, pitch correlation, percentage of Dolch (high-frequency) words with minimal latency, and change rate of normalized production correlation. n.

Table 11 shows the adjusted R^2 of successive models that added each of these features in stepwise linear regression.

Table 11: Top predictors in correlation+latency model

Posttest fluency		Posttest comprehension	
Feature	Adjusted R^2	Feature	Adjusted R^2
Percentage of words accepted as read fluently	0.836	Percentage of words with defined latency	0.619
Normalized production time	0.845	Correlation of latency	0.652
Speedup in normalized production	0.859	Correlation of pitch	0.674
Correlation of latency	0.860	Percentage of Dolch words	0.673

In comparison, Miller and Schwanenflugel [2008] 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. However, latency predicted fluency, and pitch correlation predicted comprehension, rather than vice versa. Our results make sense insofar as disfluency – slow, laborious reading with many false starts and restarts – is closely related to latency. Also, one expects intonation to reflect understanding as it requires sensitivity to syntax, contrastive stress, and text style.

4.3 Comparing to rubric-based prediction

How did our automatically computed template-based features compare to the hand-scored fluency rubric in predicting the same students' test scores? To find out, we trained models on the hand-scored rubric ratings for the 10 students using Weka's linear regression with the M5 feature selection method. We evaluated the trained models using leave-1-out cross-validation to avoid overfitting. Table 12 summarizes rubric-based prediction of test scores.

Table 12: Experiments in rubric-based prediction of test scores

Model	Fluency rubric
Input	Sentence level rubric scores on 4 dimensions of 200 sentences read by 10 students
Train	WEKA linear regression with M5 feature selection
Aggregate	Average over 20 sentences for each student
Predict	Posttest fluency and comprehension scores
Evaluate	Cross-validated correlation of actual and predicted values

We fit template-based, pretest-based, and pretest+template models to the 45 remaining students' posttest scores. To evaluate these models, we used the data from the 10 students with rubric scores as a held-out test set. This procedure enabled us to compare the rubric-based and other models on the data from the same 10 students.

Table 13 compares correlations for the four types of model. All models predicted posttest fluency with correlation 0.8 or better, thanks to the heterogeneity of a small sample of students stratified by fluency. In predicting posttest comprehension, the template-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 13: Correlation of predicted to actual posttest scores for 10 rubric-scored students

Model	Posttest fluency	Posttest comprehension
Rubric-based	0.930	0.300
Pretest-based	0.946	0.786
Template-based	0.946	0.454
Pretest + template	0.801	0.310

4.4 Varying 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 template-based models trained on the same 10 students. One training condition used only the 200 sentences scored by the rubric annotator, but instead of their manual scores it used their automated features. Another condition trained on all the sentences read by these 10 students (2503 sentences, ranging from 95 to 464 sentences per student), not just the hand-scored ones.

Table 14 compares the leave-out-one cross-validated accuracy of both these models to the models evaluated in Table 13. Trained on the same 200 sentences, the template-based model explained 3% less of the variance in fluency than the rubric-based model, but 4% more of the variance in comprehension.

Table 14: Cross-validated correlations of predicted to actual posttest scores using template models trained on different amounts of data

Training data for template models	Posttest fluency	Posttest comprehension
Trained on only 200 sentences from 10 students	0.902	0.342
Trained on all sentences from 10 students	0.981	0.713
Trained on all 45 other students	0.946	0.454

As expected, using more data from the same 10 students resulted in better template-based models for fluency (explaining 8% more variance) and especially for comprehension (doubling the amount of variance explained).

The template-based models trained on the other 45 students did better than the rubric- and template-based models trained on only 200 sentences from the 10 students, but not as well as models trained on all of their data. Given enough data about the 10 students, the trained models evidently exploited some characteristics of this group, even though they were cross-validated across students to keep them from using information about the held-out student.

In short, the template-based models outpredicted the rubric-based models primarily thanks to training on more data about the 10 students than to its features, which were actually a bit worse than the rubric-based features at predicting fluency from the same data, and only slightly better than them at predicting comprehension. The rest of this paper does not depend on human-scored ratings of oral reading prosody.

5. GENERALIZED MODELS

Section 4.2 showed (in Table 10) that correlating children's oral reading prosody against adult narrations of the same text helped (along with other sentence features in Table 3) to predict their fluency and comprehension posttest scores about as well as or better than pretest scores did, and predicted their gains considerably better. In this section we tackle the same problem of assessing children's oral reading prosody, but eliminate the need

for an adult narration of each sentence scored. Instead, we adapt prior work [Jurafsky and Martin, 2008] that trains models of duration, F0, and intensity in order to map text to prosody. We train similar models, but instead of using them to prescribe a specific prosodic contour, we use them to assess children's prosody. We train our model on multiple adult voices so it is not specific to the idiosyncrasies of one speaker. This model also lets us score readings of new text unnarrated by adults.

Prosody can be quantified by duration, pitch, and intensity. As Section 4.2 reported, we found duration-based features strongest in predicting paper tests of fluency and comprehension. Our work on detecting prosody improvement [Duong and Mostow, 2009] also found duration-based features strongest. This section therefore focuses on duration: how we train a model to synthesize durations; the features it uses; and how we transform it into a model to assess oral reading prosody.

5.1 Duration model for synthesis

Several duration models, either rule-based or statistical, have been shown to work well in speech synthesis. Most well-known among the rule-based methods is the method of Klatt [1979], which uses rules to model the average duration of a phone given its surrounding context. Examples of good machine learning methods for prosody models are decision trees [Breiman et al., 1984; Quinlan, 1986] and the sum-of-products model [van Santen, 1994; van Santen, 1997; van Santen, 1998]. Using adult narrations of hundreds of stories in the Reading Tutor as training data, we decided to train a decision tree model of phone duration, using tools in the Festival Speech Synthesis System [Hunt and Black, 1996]. Given recorded, transcribed utterances, the trainer computes a set of features for each phone and builds a decision tree using these features. Rather than simply using all features, the trainer uses a greedy stepwise approach to select which set of features to use. Each step tests the features to find the best feature to add next. Given a new text to synthesize, the model generates each phone's duration as follows. First it computes the selected features of the phone and its surrounding context, up to the utterance level, to place the phone in the appropriate leaf node. It then uses the mean duration of all training data instances that have fallen into that leaf node as the synthesized duration for the phone. We now describe the features used in the model.

5.2 Features in duration model

Our duration model uses features of the phone itself, as well as contextual features about the syllable structure and the word containing the phone. Phone level features include the phone name, its position in the syllable, whether it occurs before or after the syllable's nucleus, and whether it's a consonant or a vowel. For a vowel, we compute length (short, long, diphthong or schwa), height (high, mid, or low), frontness (front, mid, or back), and roundedness (rounded or unrounded). For a consonant, we include type (stop, fricative, affricative, nasal, lateral, or approximant), place of articulation (labial, alveolar, palatal, labio-dental, dental, velar, or glottal) and voicing

(voiced or unvoiced). For some of these features, we also look two phones to the left and right. Syllable level features include number of phones in onset and coda, position in word, distance to end of the phrase, number of syllables from last and next phrase breaks, number of stressed syllables from last and next phrase breaks, and the lexical stress of the syllable. At the word level, we use the number of syllables in the word, and its context-sensitive part of speech.

5.3 Using a synthesis model for assessment

So far we have simply described common practice in prosodic synthesis. The novelty of our method is in transforming a model for synthesizing speech into a model for assessing children's oral reading. To this end, we train the model similarly to what is done for synthesis, but use it differently. Instead of using the mean duration at each leaf node of the decision tree to prescribe the duration for a phone being synthesized, we use the mean (and possibly the standard deviation) of all the instances at the leaf node to score the observed duration of the phone being assessed.

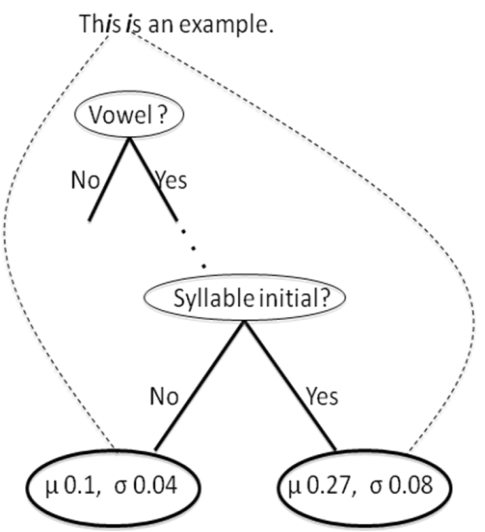


Figure 3: Decision tree fragment with two /IH/ nodes

Figure 3 depicts a fragment of a simple (made-up) tree covering the two instances of the phone /IH/ in speaking the sentence *Th<u>i</u>s <u>i</u>s an example*. The /IH/ in *Th<u>i</u>s* is in the middle of the syllable, so when it reaches the "Syllable initial?" test, it follows the left branch to the bottom left leaf. Conversely, the /IH/ in *<u>i</u>s* occurs at the beginning of the syllable, so it follows the right branch to the bottom right leaf. Each phone contributes to the training instances used to estimate the mean and standard deviation at its leaf node. According to these (fictitious) estimates, adult narrators tend to pronounce /IH/ longer in *<u>i</u>s* than in *Th<u>i</u>s*, but with greater variability. How can we use such statistics to score children's oral reading?

Given a child's utterance and a trained decision tree, we evaluated different formulas for how to compute phone-level scores and aggregate them

into an overall score for the utterance. In the following equations, u is the utterance to score, containing n phones; the p_i's are phones in that utterance; d_i is the actual duration produced by the child for phone p_i; and μ_i and σ_i are the mean and standard deviation of the training data for that phone.

1. Average log likelihood:

$$avg\_ll = \frac{1}{n}\sum_{p_i \in u} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(d_i-\mu_i)^2}{2\sigma_i^2})$$

2. Average z-score:

$$avg\_zscore = \frac{1}{n}\sum_{p_i \in u} \frac{|d_i-\mu_i|}{\sigma_i}$$

3. Pearson correlation coefficient:

$$correlation = \frac{1}{n-1}\sum_{p_i \in u} (\frac{d_i-\overline{d}}{s_d})(\frac{\mu_i-\overline{\mu}}{s_\mu})$$

where s_d and s_μ are the standard deviation of the actual duration and mean duration, respectively, of the phones in this utterance.

4. Root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{p_i \in u}(d_i-\mu_i)^2}$$

5. Mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{p_i \in u}|d_i-\mu_i|$$

6. Correlation of word durations:

$$word\_dur\_correl = \frac{1}{\#words-1}\sum_{word_j} (\frac{D_j-\overline{D}}{s_D})(\frac{\mathrm{M}_j-\overline{\mathrm{M}}}{s_\mathrm{M}})$$

where D_j and M_j represent the actual and prescribed durations of $word_j$, respectively.

7. Weighted average of word-level correlations:

$$avg\_word\_correl = \frac{1}{n}\sum_{word_j} \#phones_j \times word\_correl_j$$

Measures 1-5 combine the scores of individual phones (including silences) directly into an utterance level score. Measure 1 (average log likelihood) indicates how close the children's durations are to the model, and assumes that all the phones in the utterance are independent. It takes into account both the means and the standard deviations estimated from the training data. Measure 2 (z-score) also uses the standard deviations as well as the means, specifically to normalize the distance of an instance from the mean. Measures 3-5 consider only the mean durations, measuring how far the children's durations are from these means. Measures 4 (root mean squared error) and 5 (mean absolute error) consider only the difference between the model's means and the children's actual durations, whereas measure 3 (duration correlation) looks in addition at whether the two sequences follow similar contours.

Measures 6 and 7 first compute scores at the word level, and then combine these word level scores into utterance level scores. For Measure 6, we first compute the total durations of the phones in each word (including the preceding silence) – both the actual durations of the phones spoken by the child, and the durations prescribed by the synthesis model (namely, the mean durations of the training examples for the corresponding leaf nodes). We then compute the correlation between the actual and prescribed durations of the words in the utterance. For Measure 7, we first correlate the actual and prescribed durations of the phones in each word, obtaining one correlation for each word. We then take the weighted average of these correlations, using the number of phones in each word as the weight.

When we compute these scores, we must handle cases where a child misread, skipped, or repeated words. Just as for template models, we consider only the child's first attempt to read a word, and exclude words that the child skipped, or that the tutor assisted before the child could attempt them. We count the number of such words in each utterance, and use the percentage of included words to weight the score of each utterance when we combine the scores of all the utterances. This weighting scheme gives higher weights to sentences with fuller observations of the child's performance.

Our models leave room for further improvement. We plan to train similar models for latency, pitch, and intensity, in order to derive a more comprehensive set of scores for children's oral reading prosody. The Festival toolkit provides straightforward modules to train synthesis models for these features, and transforming them into models for scoring should be very similar to transforming the duration model. Another limitation consists of computing decision tree estimates (mean and standard deviation) solely from the statistics at the leaf level. Although taking standard deviation into account uses more information than synthesizing based solely on the mean, estimates at leaf nodes may be based on sparse data. To mitigate this problem, we had a rule to stop splitting the tree whenever there are fewer than 20 training instances. A more principled approach would use deleted interpolation [Magerman, 1995], which smoothes sparse estimates of leaf-level probabilities, conditioned on many features of the phone, by combining them with better-estimated but less specific probabilities at higher levels in the tree. However, before extending the approach to evaluate latency, pitch, and intensity or use more sophisticated estimates, it makes sense to evaluate the basic idea, as we now describe.

6. EVALUATION OF GENERALIZED MODELS

What is a better way to score children's oral reading prosody – comparison to fluent adult narrations of the same sentences, or a generalized model trained on those narrations? To address this question, we evaluated both methods.

We combined the utterance-level scores for each student as described in Section 5.3, to arrive at 7 different scores for the student. We then trained a linear regression model to predict students' test performance, using these scores as independent variables. Table 15 summarizes the overall evaluation procedure.

We compared our generalized models against template models. To make the comparison fair, the template models used only features related to duration, not pitch or intensity. These features included child-adult correlations for word production, latency, and reading time, both raw and after normalizing for word length. Here production or word duration is the time to speak the word; latency is the time between successive text words, including "false starts, sounding out, repetitions, and other insertions, whether spoken or silent" [Beck et al., 2004; Mostow and Aist, 1997]; and word reading time is the sum of production and latency.

Table 15: Generalized model experiments

Model	Generalized model
Data	24,816 sentence narrations by 20 adults; 77,600 sentence readings by 55 students aged 7-10 with test scores
Train	Regression trees to map text features to adult phoneme durations; linear regression (Enter and Stepwise) to predict test scores from child's prosody
Aggregate	7 formulas to combine phone level features into sentence level prosodic features
Predict	Posttest comprehension and fluency scores
Evaluate	Adjusted R^2 of predicted vs. actual scores

In our experiments, we used SPSS's linear regression function, with either the Stepwise or Enter method of selecting features. The Enter method simply includes all the features in the model, whereas the bidirectional Stepwise method inserts or removes one feature at each step based on an F-test. This greedy method sometimes does worse, so we tried both methods and reported the adjusted R^2 of the two.

Table 16 compares the adjusted R^2 of two types of models. The template models use features computed by correlating the child's and adult narrator's prosodic contour for each sentence, as described in Section 3. The generalized models use scores output by the synthesis model trained on the same set of adult narrations, as described in Section 5.

As Table 16 shows, the generalized model accounts for 7.2% more variance in fluency and 23.3% in comprehension than the template model, based on adjusted R^2 for Enter or Stepwise, whichever is higher. The fact that the generalized models out-predicted the template models means they gained more by generalizing across different sentences than they lost by ignoring the sentence details they did not capture.

Table 16. Adjusted R^2 of competing models

Dependent variable	**Template**		**Generalized**	
	Enter	**Stepwise**	**Enter**	**Stepwise**
Posttest fluency	0.555	0.565	0.635	0.637
Posttest comprehension	0.355	0.362	0.580	0.595

Table 17. Adjusted R^2 of pretest with vs. without generalized model scores

Dependent variable	Pretest		Pretest + generalized	
	Enter	Stepwise	Enter	Stepwise
Posttest fluency	0.873	0.852	0.852	0.866
Posttest comprehension	0.783	0.792	0.813	0.813

Speakers did not overlap between the adult narrations used to train the generalized models and the children's oral reading used to test them, but text overlapped: the children read the same sentences narrated by the adults. However, the amount of training data (24,816 sentences, times roughly 50 phonemes per sentence) is so large relative to the size of the phoneme-level decision tree trained on them (11,123 decision nodes) as to assuage fears of overfitting those particular sentences. Consequently the generalized models should perform well on unseen sentences drawn from a similar distribution.

Pretest scores are typically strong predictors of post-test scores, so we also tested whether generalized models with pretest score as an additional feature did better than pretest score alone. Table 17 shows that pretest scores achieved high adjusted R^2. Adding the generalized model's features actually reduced adjusted R^2 for fluency by 0.7%, from 0.873 to 0.866 (conceivably an artifact of overfitting by Enter), but accounted for 2.1% additional variance in comprehension.

In models constructed by stepwise regression, we noticed that the average log likelihood score for the utterance (Measure 1) was always the first feature to be selected, sometimes the only one. This finding demonstrates the value of incorporating the standard deviation of phone durations into the scores.

7. CONTRIBUTIONS AND FUTURE WORK

In this paper we presented two methods to assess oral reading prosody automatically, and evaluated them on a large corpus of real oral reading by real children recorded by Project LISTEN's Reading Tutor in real schools.

The template model method scores a child's oral reading prosody by correlating it against a fluent adult narration of the same sentence. More precisely, it uses such correlations as input variables in a linear regression model to predict students' test scores and gains.

The generalized model method trains a prosodic synthesis model on the same fluent adult narrations and transforms it to score children's oral reading prosody. We showed how to use the trained synthesis model to score each phone, and investigated seven different formulas to combine phone scores to score words and utterances. Two of these formulas exploit standard deviations of the feature values underlying each leaf of the trained decision tree. The method trains generalized models of oral reading prosody that could be used to score sentences not in the training corpus.

We evaluated the template method against a fluency rubric on a sample of 10 students stratified by reading proficiency. Two human judges scored 20 sentence readings by each student. However, their inter-rater reliability was low, especially for individual sentences, making them problematic to predict automatically; we did better in estimating scores at the student level.

We evaluated both of our methods indirectly by their ability to predict students' scores on fluency and comprehension tests. The generalized models beat the template models on both tasks. Moreover, the generalized model method is qualitatively superior in that it eliminates the requirement for an adult to narrate each sentence in order for the computer to score it. Although it still requires a narrated training corpus, a generalized model applies to any sentence, whereas a template model is restricted to the sentences in the corpus. The opposite result would have suggested that the phone features employed by the generalized model failed to capture enough information about the sentence text to score its prosody as accurately as comparing it to the adult narration. Evidently, the smoothness added by generalizing over multiple narrators and sentences more than compensated for whatever information was lost by ignoring sentence details unrepresented by the phone and context features in the generalized model.

Besides extending the generalized model method to score latency, pitch, and intensity, we are working to mine the trained models for useful knowledge [Mostow and Sitaram, 2011]. First, characterizing which sentences are scored more accurately by a template model than by a generalized model may reveal important text features that govern oral reading prosody but are missing from the generalized model. Second, characterizing which decision tree nodes are especially strong predictors of comprehension test scores or gains may reveal specific prosodic knowledge that facilitates reading comprehension. Third, characterizing which sentences are especially strong predictors of test scores or gains may reveal features of text that make it especially diagnostic for scoring oral reading prosody or predicting comprehension. Evidence for the potential value of this direction comes from recent work [Benjamin and Schwanenflugel, 2010] relating children's oral reading of two text passages to their comprehension scores. Their performance on the harder text passage turned out to be highly predictive, in contrast to the easier passage. This finding suggests that we might improve the predictive accuracy of our models simply by ignoring children's prosody on easy text. More generally, it suggests that discovering features of words and sentences especially predictive of test scores and gains may shed new light on how to assess, predict, and improve children's oral reading fluency and comprehension.

ACKNOWLEDGMENTS

We thank Drs. Paula Schwanenflugel and Melanie Kuhn for their expertise, the educators, students, and LISTENers who helped generate our data, and Anders Weinstein for implementing the prosodic display depicted in Figure 2. Last but definitely not least, we thank the reviewers for their constructive criticisms and helpful suggestions.

REFERENCES

BALOGH, J., BERNSTEIN, J., CHENG, J. and TOWNSHEND, B. 2007. Automatic evaluation of reading accuracy: assessing machine scores. In *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE)*, Farmington, PA, October 1-3, 2007, M. ESKENAZI, Ed., 112-115.

BANERJEE, S., BECK, J.E. and MOSTOW, J. 2003. Evaluating the effect of predicting oral reading miscues. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, Geneva, Switzerland, September 1-4, 2003, 3165-3168.

BEATTIE, V.L. 2010. Scientific Learning Reading Assistant™: CMU Sphinx technology in a commercial educational software application. In *CMU Sphinx Users and Developers Workshop*, Dallas, TX, March13, 2010.

BECK, J.E., JIA, P. and MOSTOW, J. 2004. Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning* 2(1-2), 61-81.

BECK, J.E. and MOSTOW, J. 2008. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students [Best Paper Nominee]. In *9th International Conference on Intelligent Tutoring Systems*, Montreal, June 23-27, 2008, 353-362.

BECK, J.E. and SISON, J. 2006. Using knowledge tracing in a noisy environment to measure student reading proficiencies. *International Journal of Artificial Intelligence in Education (Special Issue "Best of ITS 2004")* 16(2), 129-143.

BENJAMIN, R.G. and SCHWANENFLUGEL, P.J. 2010. Text complexity and oral reading prosody in young readers. *Reading Research Quarterly* 45(4), 388-404.

BERNINGER, V.W., ABBOTT, R.D., VERMEULEN, K. and FULTON, C.M. 2006. Paths to reading comprehension in at-risk second-grade readers. *Journal of Learning Disabilities* 39(4), 334-351.

BERNSTEIN, J., COHEN, M., MURVEIT, H., RTISCHEV, D. and WEINTRAUB, M. 1990. Automatic evaluation and training in English pronunciation. In *International Conference on Speech and Language Processing (ICSLP-90)*, Kobe, Japan, 1990, 1185-1188.

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 ICSLP*, Antwerp, Belgium, October 2007, 2007.

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

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. 1984. *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove, CA.

CMU 2008. The CMU Sphinx Group open source speech recognition engines [software at http://cmusphinx.sourceforge.net].

COMPTON, D.L. and APPLETON, A.C. 2004. Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research & Practice* 19(3), 176–184.

DENO, S.L. 1985. Curriculum-Based Measurement: The emerging alternative. *Exceptional Children* 52(3), 219-232.

DUONG, M. and MOSTOW, J. 2009. Detecting prosody improvement in oral rereading. In *Second ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, Wroxall Abbey Estate, Warwickshire, England, September 3-5, 2009, M. RUSSELL, Ed.

GEROSA, M. and NARAYANAN, S.S. 2008. Investigating automatic assessment of reading comprehension in young children. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2008*, 2008, 5057-5060.

HAGEN, A., PELLOM, B. and COLE, R. 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication* 49(12), 861-873.

HASBROUCK, J.E. and TINDAL, G.A. 2006. Oral reading fluency norms: a valuable assessment tool for reading teachers. *The Reading Teacher* 59(7), 636-644.

HUNT, A. and BLACK, A. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP 96*, Atlanta, GA, 1996, 373-376.

JURAFSKY, D. and MARTIN, J.H. 2008. *Speech and Language Processing (2nd ed.)*. Pearson Prentice Hall, Upper Saddle River, NJ.

KLATT, D.H. 1979. Synthesis by rule of segmental durations in English sentences. In *Frontiers of Speech Communication Research*, B.E.F. LINDBLOM and S. OHMAN, Eds. Academic, 287-299.

KUHN, M.R., SCHWANENFLUGEL, P.J., MORRIS, R.D., MORROW, L.M., BRADLEY, B.A., MEISINGER, E., WOO, D. and STAHL, S.A. 2006. Teaching children to become fluent and automatic readers. *Journal of Literacy Research* 38(4), 357-387.

KUHN, M.R. and STAHL, S.A. 2003. Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology* 95(1), 3–21.

LEE, K., HAGEN, A., ROMANYSHYN, N., MARTIN, S. and PELLOM, B. 2004. Analysis and detection of reading miscues for interactive literacy tutors. In *20th International Conference on Computational Linguistics (Coling)*, Geneva, Switzerland, August, 2004.

MAGERMAN, D.M. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, MIT, Cambridge, MA, June 26-30, 1995, 276-283.

MILLER, J. and SCHWANENFLUGEL, P.J. 2006. Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology* 98(4), 839-853.

MILLER, J. and SCHWANENFLUGEL, P.J. 2008. A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly* 43(4), 336–354.

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

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. *Journal of Educational Computing Research* 29(1), 61-117.

MOSTOW, J. and BECK, J. 2005. Micro-analysis of fluency gains in a Reading Tutor that listens: Wide vs. repeated guided oral reading, Twelfth Annual Meeting of the Society for the Scientific Study of Reading, Toronto.

MOSTOW, J., BECK, J., BEY, J., CUNEO, A., SISON, J., TOBIN, B. and VALERI, J. 2004. Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning* 2, 97-134.

MOSTOW, J., BECK, J., CUNEO, A., GOUVEA, E., HEINER, C. and JUAREZ, O. 2010. Lessons from Project LISTEN's Session Browser. In *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, S.R. VIOLA, M. PECHENIZKIY and R.S.J.D. BAKER, Eds. CRC Press, Taylor & Francis Group, New York, 389-416.

MOSTOW, J., ROTH, S.F., HAUPTMANN, A.G. and KANE, M. 1994. A prototype reading coach that listens [AAAI-94 Outstanding Paper]. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Seattle, WA, August, 1994, American Association for Artificial Intelligence, 785-792.

MOSTOW, J. and SITARAM, S. 2011. Mining data from Project LISTEN's Reading Tutor to analyze development of children's oral reading prosody. In: D. COMPTON (Editor), Eighteenth Annual Meeting of the Society for the Scientific Study of Reading, St. Pete Beach, Florida.

NRP 2000. 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. 00-4769, National Institute of Child Health & Human Development. At www.nichd.nih.gov/publications/nrppubskey.cfm, Washington, DC.

O'CONNOR, R.E., WHITE, A. and SWANSON, H.L. 2007. Repeated reading versus continuous reading: Influences on reading fluency and comprehension. *Exceptional Children* 74(1), 31-46.

PINNELL, G.S., PIKULSKI, J.J., WIXSON, K.K., CAMPBELL, J.R., GOUGH, P.B. and BEATTY, A.S. 1995. Listening to children read aloud: Oral reading fluency, National Center for Educational Statistics, Washington, DC.

PRICE, P., TEPPERMAN, J., ISELI, M., DUONG, T., BLACK, M., WANG, S., BOSCARDIN, C.K., HERITAGE, M., PEARSON, P.D., NARAYANAN, S. and ALWAN, A. 2009. Assessment of emerging reading skills in young native speakers and language learners. *Speech Communication* 51, 968–984.

QUINLAN, J.R. 1986. Induction of decision trees. *Machine Learning* 1, 81-106.

RASINSKI, T., RIKLI, A. and JOHNSTON, S. 2009. Reading fluency: More than automaticity? More than a concern for the primary grades? *Literacy Research and Instruction* 48, 350-361.

SCHWANENFLUGEL, P.J., HAMILTON, A.M., KUHN, M.R., WISENBAKER, J.M. and STAHL, S.A. 2004. Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology* 96(1), 119-129.

SCHWANENFLUGEL, P.J., MEISINGER, E.B., WISENBAKER, J.M., KUHN, M.R., STRAUSS, G.P. and MORRIS, R.D. 2006. Becoming a fluent and automatic reader in the early elementary school years. *Reading Research Quarterly* 41(4), 496-522.

SHROUT, P.E. and FLEISS, J.L. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86(2), 420-428.

SOPER, D.S. 2010. The Free Statistics Calculators Website. http://www.danielsoper.com/statcalc/.

TEPPERMAN, J., KAZEMZADEH, A. and NARAYANAN, S. 2007. A text-free approach to assessing nonnative intonation. In *Proceedings of ICSLP*, Antwerp, Belgium, October 2007, 2007.

TEPPERMAN, J., LEE, S., ALWAN, A. and NARAYANAN, S.S. 2011. A generative student model for scoring word reading skills. *IEEE Transactions on Audio, Speech, and Language Processing* 19(2), 348-360.

VAN SANTEN, J.P.H. 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language* 8, 95-128.

VAN SANTEN, J.P.H. 1997. Segmental duration and speech timing. In *Computing Prosody: Computational Models for Processing Spontaneous Speech*, Y. SAGISAKA, N. CAMBELL and N. HIGUCHI, Eds. Springer, 225-250.

VAN SANTEN, J.P.H. 1998. Timing. In *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*, R. SPROAT, Ed. Kluwer, 115-140.

WOODCOCK, R.W. 1998. *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. American Guidance Service, Circle Pines, Minnesota.

ZHANG, X., MOSTOW, J. and BECK, J.E. 2007. Can a computer listen for fluctuations in reading comprehension? In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, Marina del Rey, CA, July 9-13, 2007, R. LUCKIN, K.R. KOEDINGER and J. GREER, Eds. IOS Press, 495-502.

ZUTELL, J. and RASINSKI, T.V. 1991. Training teachers to attend to their students' oral reading fluency. *Theory into Practice* 30(3), 211-17.