ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Detecting Prosody Improvement in Oral Rereading

Minh Duong *Jack Mostow*

Project LISTEN, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA, USA
mnduong@cs.cmu.edu mostow@cs.cmu.edu

Abstract

A reading tutor that listens to children read aloud should be able to detect fluency growth – not only in oral reading rate, but also in prosody. How sensitive can such detection be? We present an approach to detecting improved oral reading prosody in rereading a given text. We evaluate our method on data from 133 students ages 7-10 who used Project LISTEN's Reading Tutor. We compare the sensitivity of our extracted features in detecting improvements. We use them to compare the magnitude of recency and learning effects. We find that features computed by correlating the student's prosodic contours with those of an adult narration of the same text are generally not as sensitive to gains as features based solely on the student's speech. We also find that rereadings on the same day show greater improvement than those on later days: statistically reliable recency effects are almost twice as strong as learning effects for the same features.

1. Introduction

An important task of any intelligent tutor is to detect improvement in the skills it is trying to help students learn. In the case of oral reading fluency, it is important to distinguish between reading a new text and rereading a text the student has seen before, whether earlier the same day or less recently. The ultimate goal of fluency practice and instruction is skilled reading of new text. However, repeated reading of the same text is a popular and effective form of fluency practice [1-4], thanks in part to the presumed motivational value of giving students feedback on their improvement in reading speed.

We recently [5] developed a method to assess oral reading prosody automatically based on several features, and evaluated it on children's reading of new text in Project LISTEN's Reading Tutor, both by comparing to human scoring, and by predicting test scores and gains in fluency and comprehension.

Here we apply the same assessment method to rereading:
Section 2 summarizes relevant aspects of the Reading Tutor.
Section 3 describes prosodic features of oral reading.
Section 4 measures the relative sensitivity of those features.
Section 5 compares the size of recency and learning effects.
Section 6 tests whether any features are sensitive enough for individual students' improvements to be statistically reliable.
Section 7 outlines contributions, limitations, and future work.

2. Project LISTEN's Reading Tutor

Our data consist of assisted oral reading recorded by Project LISTEN's Reading Tutor, which listens to children read aloud, and helps them learn to read [6]. 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) [7] to listen to the child read the sentence aloud, to track the child's position in the text to detect deviations from it, and to identify the start and end points of each word and silence in the recorded oral reading [8]. 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. The spoken feedback uses a time-aligned recording of each sentence by an adult narrator.

2.1. Data set

The data for this paper came from children in grades 2-4 (ages 7-10) who used the Reading Tutor during the 2005-2006 school year. They read a total of 77,693 sentences, including 4,901 distinct sentences that they reread a total of 29,794 times, averaging 1.66 rereadings per reread sentence, with a median of 1. Our 164 students ranged from 1 to 1891 data points (pairs of successive readings of sentences). To reduce the number of outliers, we excluded students with fewer than 5 data points, which left us with 133 students.

3. Features of oral reading prosody

To measure students' oral reading prosody, we extract various features from each sentence of their recorded speech. We use the same features that were explored in previous work [5]. They are of two types: raw features and correlational features.

Raw features are based solely on the student's speech. They include average word production time, average inter-word latency, and average word reading time, which is the sum of production and latency. Inter-word latency (or simply 'latency') is the time that elapses between reading successive text words, including "false starts, sounding out, repetitions, and other insertions, whether spoken or silent" [9, 10]. We normalize these time features by word length, yielding three more features. The last raw feature is pause frequency, which measures how often a student pauses for more than 10 ms before a word, or the Reading Tutor's ASR rejects a word as read incorrectly.

Correlational features are inspired by previous analyses of children's oral reading prosody by Schwanenflugel and colleagues, 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 [11-13]. Each correlational feature is computed as the Pearson correlation coefficient between the prosodic contour of a student and that of the adult narration. We compute correlations for word production time, latency, and word reading time (both unnormalized and normalized), as well as the mean fundamental frequency and intensity for each word. F0 and

intensity are computed using the Praat pitch tracker [14]. Finally, we also include in this group the sentence's pitch variation, computed as the standard deviation of the words' fundamental frequencies. It is not based on any correlation, but shares a common characteristic with all other correlational features: a higher value indicates more fluent, expressive reading.

3.1. Descriptive statistics

Here we report some descriptive statistics about the original data set (before excluding outliers). To analyze improvement in rereading the same text, we considered only sentences the student saw more than once. For each such sentence, we computed the gain between successive values of each feature as the *decrease* from the prior value of a raw feature or the *increase* from the prior value of a correlational feature, so that positive gains all represent improvement. For each feature, we computed each student's mean feature value, mean gain, and relative gain (mean gain / mean prior value). Table 1 shows the medians of these means. We use medians to counteract distortion of means by outliers or small, noisy samples. (Using medians at the individual level would have lost too much information about features where over half of the values are the same.)

Table 1: Descriptive statistics of features

Feature	Median of mean value	Median of mean gain	Median relative gain (%)
avg_duration (s)	0.552	0.030	5.2
avg_norm_duration (s)	0.137	0.007	5.6
avg_norm_production (s)	0.112	0.005	4.4
pause_frequency (%)	27.2	2.2	10.5
avg_production (s)	0.467	0.017	3.4
avg_norm_latency (s)	0.026	0.002	12.4
avg_latency (s)	0.094	0.010	12.9
correl_duration	0.496	0.022	3.2
correl_production	0.619	0.020	3.1
correl_latency	0.430	0.012	3.0
correl_norm_latency	0.354	0.019	3.5
correl_norm_duration	0.284	0.014	1.6
pitch_variation (Hz)	40.451	0.604	1.6
correl_norm_production	0.342	0.010	1.4
correl_pitch	0.134	-0.001	-10.2
correl_intensity	0.238	-0.003	-5.9

As Table 1 shows, words took a median of 0.094 seconds latency plus 0.467 seconds to say the word, totaling 0.552 seconds, or 0.137 seconds per letter. The pause_frequency of 27.2% reflects disfluent reading. The features with the highest relative gains (> 10%) were avg_latency, avg_norm_latency, and pause_frequency, indicating successively fewer and shorter pauses as fluency increased. Children's production, duration, and latency correlated the most strongly with the adult narrators, and pitch and intensity the least.

4. Compare sensitivity of features

Here we present comparisons between features. For each feature and student, we had a sample of change values. We computed Cohen's d effect size for each individual student as the mean of this sample divided by its standard deviation [15]. We quantified the sensitivity of each feature as the median of these per-student individual effect sizes.

The "Any day" column in Table 2 lists the median effect size of each feature, in decreasing order. (Section 5 explains the "Same day" and "Later day" columns.) From Table 2, we observe that all raw features have larger effect sizes than correlational features. Why? Correlational features are computed by correlation between a student's prosodic contours and those of the adult narrations of the same sentences. Their values, therefore, depend not only on the student's speech but also on the particular adult's, a source of additional variability and measurement noise. Correlational features are useful for assessing the *quality* of students' prosody, since better readers read more like adults [13]. But apparently raw features are more sensitive to its *improvement*.

Table 2: Median effect sizes for different features

ID	Feature	Any day	Same day	Later day
1	avg_duration	0.138	**0.218**	0.130
2	avg_norm_duration	0.133	**0.222**	0.138
3	avg_norm_production	0.127	**0.232**	0.117
4	pause_frequency	0.101	**0.146**	0.095
5	avg_production	0.097	**0.187**	0.084
6	avg_norm_latency	0.084	0.133	0.083
7	avg_latency	0.077	0.124	0.086
8	correl_duration	0.050	0.034	0.061
9	correl_production	0.049	0.045	0.046
10	correl_latency	0.034	0.042	0.036
11	correl_norm_latency	0.034	0.061	0.027
12	correl_norm_duration	0.023	0.018	0.024
13	pitch_variation	0.019	-0.015	0.012
14	correl_norm_production	0.015	-0.034	0.021
15	correl_pitch	-0.001	-0.001	-0.011
16	correl_intensity	-0.016	-0.009	-0.015

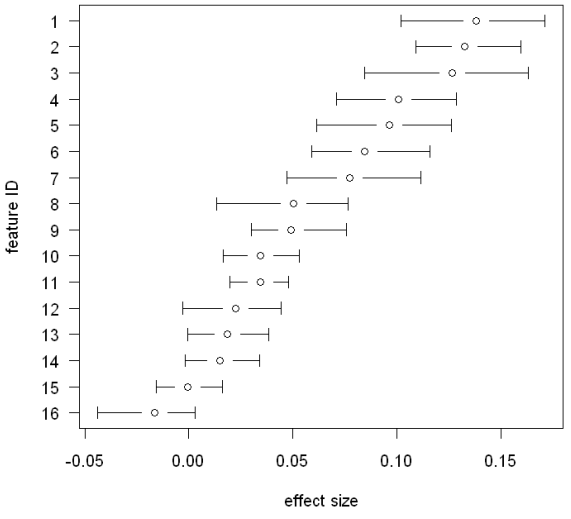

Figure 1: Confidence intervals of median effect sizes

To analyze the reliability of the median effect sizes, we computed confidence intervals around them, using the bootstrapping method [16]. Assuming normality among the medians of the 100 samples that we generated with R's

boot::boot function, we computed a 95% confidence interval around each feature's median. (Without the normality assumption, some of the confidence intervals aren't quite as tight.) Figure 1 shows the resulting confidence intervals. It shows, for example, that the median effect size (0.138) for avg_duration (feature 1 in Table 2) is reliably greater than for features 8-16. Improvement in this sentence feature corresponds to reduction in the time to read the sentence. Thus speed-up in oral reading rate is the most sensitive indicator of improvement, which is consistent with the widespread use of timed oral (re-)reading for that purpose. The small effect size reflects the subtlety of the effect – a reduction of a few milliseconds in average word reading time.

Conversely, correl_intensity (feature 16) is less sensitive to improvement than any other feature, significantly less so for all but correl_pitch, and neither of their effect sizes differs significantly from 0. However, effect sizes for most of the other features are significantly greater than 0, though some of them are very small.

5. Compare learning vs. recency effects

When a student reads the same text twice in a row it is not surprising if reading is more fluent the second time around, due to short-term memory recency effects. It is more impressive when the improved rereading occurs on a later day, because it indicates learning and retention.

To compare the relative magnitudes of these two effects, we disaggregated our analysis of improvement in reading a sentence into two cases. In one case, the rereading occurred on the same day as the previous reading. In the other case, the rereading occurred on a later day. For both cases, we compute each feature's median effect size as described in Section 5.

Table 2 compares these two effect sizes side by side in the "Same day" and "Later day" columns. Most of them are higher for same-day rereading than for later-day rereading.

To see which such differences are statistically reliable, we did a one-tailed, paired t-test for each feature, pairing each student's effect sizes for the two cases. Table 2 indicates the results of these tests by **boldfacing** median effect sizes that are significantly greater (at the 0.05 level) than for the other case.

The results confirm our hypothesis that same-day rereading should show greater improvement than later-day rereading due to recency effects. Moreover, it quantifies the difference. For every feature where the difference is significant, same-day improvement is higher than later-day improvement – by a ratio of 1.8, on average.

6. Are individual improvements reliable?

The statistical significance tests described above are for overall differences between features or same- vs. later-day. What about for individuals? Are any features sensitive enough to detect improvements that are statistically reliable for particular students? To answer this question, we conducted a paired t-test for each feature and student to evaluate the significance of the hypothesis that the student's mean gain exceeds 0. For each feature, Table 3 reports the percentage of students whose individual improvement in that feature was statistically significant at the 0.05 level. As before, the "Any day," "Same day," and "Later day" columns shows results for all rereadings, same-day rereadings, and later-day rereadings.

We *italicize* percentages that fell below 5% to indicate that they are likely due to chance, because in any random set of samples, the expected number of samples having p-value smaller than 0.05 is 5%. Table 3 lists features in the same order as in Table 2, from largest to smallest median effect size. As Table 3 shows, the raw features, which were shown to have greater effect sizes in Table 2, also have higher percentages of students with significant improvement. There are some slight changes in the ranking of the features, but in general, features with greater median effect sizes also have higher percentages of students with significant improvement.

The largest such percentage was 38.3. The actual number of students whose reading improved was much closer to 100% -- presumably about the same as the percentage of students with positive improvements whether significant or not. The reason for analyzing individual reliability is that it is important to know not only whether a class is progressing overall, but whether each individual student is improving or not, and this determination should be statistically reliable. The point of the analysis is to see how often we can meet that standard.

Table 3: % of students with significant improvement

ID	Feature	Any day	Same day	Later day
1	avg_duration	33.8	38.3	28.6
2	avg_norm_duration	38.3	35.8	32.8
3	avg_norm_production	34.6	33.3	30.3
4	pause_frequency	29.5	25.9	27.1
5	avg_production	30.8	34.6	25.2
6	avg_norm_latency	25.6	19.8	24.4
7	avg_latency	25.8	22.2	28.0
8	correl_duration	13.9	10.5	9.6
9	correl_production	9.8	5.3	12.3
10	correl_latency	3.4	4.3	4.6
11	correl_norm_latency	8.3	9.6	7.1
12	correl_norm_duration	9.0	5.3	10.5
13	pitch_variation	8.3	7.4	9.2
14	*correl_norm_production*	7.4	*1.3*	7.0
15	*correl_pitch*	*2.6*	5.5	*2.8*
16	*correl_intensity*	*3.5*	*4.1*	*2.8*

7. Contributions, limitations, and future work

In this paper, we have presented a method for detecting prosody improvement in rereading a given text and applied it to a data set of recorded speech from 133 students. We computed the improvement of each feature value in successive readings of the same sentence and computed its effect size for each student. We used each feature's median effect size as an index of its sensitivity. Our experiments showed that raw features were more sensitive to growth than correlational features. We also found that same-day rereadings exhibited nearly twice as much improvement as later-day rereading – that is, recency effects were almost twice as strong as learning effects that persist overnight or longer.

A limitation of this work is the lack of a "gold standard". We attempted to detect improvement, and succeeded in doing so, but did not verify this detected gain by any other methods. The difficulty arises from the lack of a paper test that reliably measures prosody improvement. Previous work [5] used a laboriously human-scored rubric [17] to assess the prosody of

each sentence being read. However, the same work also showed that the rubric was not reliable at the sentence level.

The previous work [5] assessed oral reading prosody for "cold reads" (first encounters) of sentences. The work reported here is complementary to it in two respects: it focuses on rereading instead of cold reading, and it moves beyond assessing prosody to detecting improved prosody. The advantage of using rereading to detect improvement is that comparing successive readings of the same sentence controls for text differences that affect oral reading prosody, thereby eliminating a major source of variance in the data we use to compare sensitivity of different features of prosody. One next step is to generalize from detecting improvement in rereading the same text to detecting improvement in reading new text. Another limitation of the current work is its reliance on adult narrations to evaluate against adult prosody. This requirement exploits a resource that happened to exist for the Reading Tutor's current text, but introduces variance among different narrators, and prevents assessment of oral reading prosody on novel text. To address both limitations, we are working to generalize the correlational features into a normative model of prosody induced from our corpus of narrations, so that given a new text, we can predict how it should be read – not in the sense of prescribing a single prosodic contour as a speech synthesizer must do [e.g., 18], but in the sense of evaluating a given prosodic contour by estimating the likelihood that a skilled narrator would produce it.

Acknowledgements

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.

References (ours at www.cs.cmu.edu/~listen)

[1] 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.

[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] Dowhower, S.L. Repeated reading revisited: Research into practice. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 1994. *10*(4): p. 343-358.

[4] 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 www.nichd.nih.gov/publications/nrppubskey.cfm: Washington, DC.

[5] Mostow, J. and M. Duong. Automated Assessment of Oral Reading Prosody. *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED2009)*, 189-196. 2009. Brighton, UK: IOS Press.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] The Incorporation of Effect Size in Information Technology, L., and Performance Research. *ITL&P Journal*, 2003. *21*(1).

[16] Efron, B. Bootstrap Methods: Another Look At The Jackknife. *The Annals of Statistics*, 1979. *7*(1): p. 1-26.

[17] 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.

[18] Jokisch, O., H. Kruschke, and R. Hoffmann. Prosodic Reading Style Simulation for Text-to-Speech Synthesis. *First International Conference on Affective Computing and Intelligent Interaction*, 426-432. 2005. Beijing, China: Springer-Verlag.