# Mining Data from Project LISTEN's Reading Tutor to Analyze Development of Children's Oral Reading Prosody

**Sunayana Sitaram and Jack Mostow**

Project LISTEN (www.cs.cmu.edu/~listen), School of Computer Science, Carnegie Mellon University
RI-NSH 4103, 5000 Forbes Avenue, Pittsburgh, PA 15213-3891
ssitaram@cs.cmu.edu, mostow@cs.cmu.edu

## Abstract

Reading tutors can provide an unprecedented opportunity to collect and analyze large amounts of data for understanding how students learn. We trained models of oral reading prosody (pitch, intensity, and duration) on a corpus of narrations of 4558 sentences by 11 fluent adults. We used these models to evaluate the oral reading prosody of 85,209 sentences read by 55 children (mostly) 7-10 years old who used Project LISTEN's Reading Tutor during the 2005-2006 school year. We mined the resulting data to pinpoint the specific common syntactic and lexical features of text that children scored best and worst on. These features predict their fluency and comprehension test scores and gains better than previous models. Focusing on these features may help human or automated tutors improve children's fluency and comprehension more effectively.

## 1. Introduction

By logging fine-grained, comprehensive records of their interactions with students, Intelligent Tutoring systems make it possible to analyze students' evolving performance in detail, whether in longitudinal studies of individual development, or in cross-sectional studies of students at different levels. Such studies offer an unprecedented opportunity to mine such data for discoveries about learning.

This paper describes one such effort to mine data collected by Project LISTEN's Reading Tutor in the 2005-06 school year. The Reading Tutor listens to children read aloud, by using Automatic Speech Recognition to track and evaluate their oral reading (Mostow et al., 2003). Here, we analyze children's development of expressive oral reading by comparing children of differing proficiency to fluent adults. We identify which features children do best and worst on, and analyze how well these features predict standard measures of reading skill and growth.

Fluent reading is defined as reading text with "speed, accuracy, and proper expression" (NRP, 2000). Oral reading rate (words read correctly per minute) is a common measure of fluency and correlates with comprehension scores, especially in early grades (Deno, 1985; Hasbrouck and Tindal, 2006), but fails to capture expressiveness. Expressiveness is the ability to read "with appropriate expression or intonation coupled with phrasing that allows for the maintenance of meaning" (Kuhn, Schwanenflugel, and Meisinger, 2010, p. 233). Expressive prosody includes

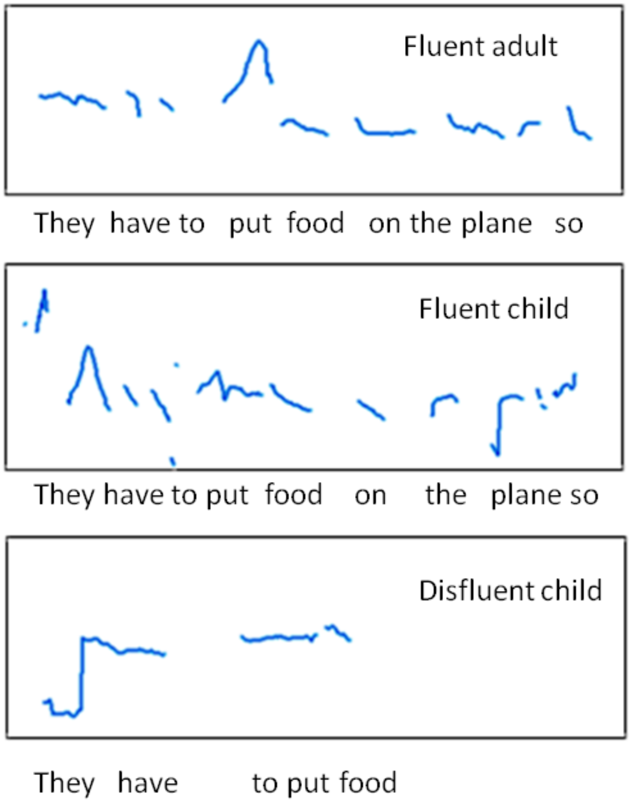

**Figure 1: Pitch contours computed in Praat (Boersma and Weenink, 2008) for the first few seconds of a fluent adult, fluent child, and disfluent child reading the sentence "They have to put food on the plane so that we can eat"**

appropriate phrasing, pause structures, stress, and rise and fall patterns (Schwanenflugel et al., 2004).

Figure 1 illustrates the contrast between a fluent and a disfluent child's pitch contours and an adult narration of the same sentence.

The rest of the paper is organized as follows. Section 2 relates this work to prior work. Section 3 details the approach we took to build and mine the prosody models. Section 4 presents the results of scoring features and using them to predict fluency and comprehension scores and gains. Section 5 concludes and discusses future work.

## 2. Relation to Prior Work

Prior work on evaluating children's oral reading prosody has been 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, 2008; Schwanenflugel et al., 2006) analyzed adult and child readings of the same short text by hand-aligning the text to a spectrogram of each reading. Given the tediousness of this operation, they computed the mean pitch of the vocalic nucleus of each word for just the first three sentences of the passage. Averaging these values across 34 adults yielded a profile of expressive reading. Correlating this profile against the corresponding values for each child quantified the expressiveness of the child's oral reading, and its changes from the end of grade 1 to the end of grade 2, so as to relate them to scores and gains on reading tests administered at those points. F0 (fundamental frequency or pitch) match and the number of pausal intrusions were the best indicators of prosodic change between first and second grades.

Mostow and Duong (2009) scaled up this manual analysis of oral reading prosody by using the Reading Tutor's ASR-based time alignment of text to oral reading. They computed prosodic contours of thousands of sentences read by children over the course of a semester. They used the Reading Tutor's single adult narration of a sentence instead of multiple adult narrations. They computed contours for latency, duration and intensity in addition to pitch. They also used a somewhat different set of features from the ones used by Schwanenflugel et al. (2006) or Benjamin and Schwanenflugel (2010). Mostow and Duong (2009) compared a child's single utterance to the same utterance spoken by a fluent adult narrator, which they called the template model. They validated the approach by predicting fluency and comprehension scores and gains, and outperformed a human-based rubric.

Duong and Mostow (2010) used a generalized model of duration trained on a corpus of adult narrations and scored children based on these models. The generalized models predicted fluency and comprehension scores and gains better than the template model but not better than pretest alone. Combining pretest scores with the generalized model predicted scores and gains better than pretest alone (Duong, Mostow, and Sitaram, 2011).

Duong and Mostow (2009) detected improvement in prosody on successive readings of the same sentence but found that features based on correlating children's prosodic contours to adult contours were not as sensitive in detecting improvement as other features involving children's speech. Here, we generalize this approach to detecting improvement on features and identifying which features best predict fluency and comprehension gains and scores (Miller and Schwanenflugel, 2008)

Miller and Schwanenflugel (2006) describe a study in which 80 third grade children read a passage containing three of the following types of sentences – basic declaratives, basic quotatives, wh- questions, yes–no questions, complex adjectival phrases, and phrase-final commas. The authors measured pitch and duration of pauses and regressed against the children's reading skill. More skilled readers were found to make shorter pauses during reading, somewhat less likely to pause at commas than less skilled readers, and more likely to mark sentence-final features with discernable changes in pitch.

Previous work focused on scoring children's oral reading prosody by comparing it to models of adult prosody and validating our approach by predicting fluency and comprehension test scores and gains. In this work, we mine trained models of adult prosody to identify lexical, syntactic and prosodic features of fluent reading.

## 3. Approach

Duong and Mostow (2010) built a generalized model for phoneme durations using the Festival Speech Synthesis Toolkit (Black, Taylor, and Caley, 1996-1999). The model used transcribed recordings of thousands of narrations by fluent adults to compute a set of features for each phoneme and build a decision tree using those features. In speech synthesis, the leaf node of the decision tree is used to predict the duration of a new phoneme to be synthesized, given its features. Instead of using the duration at each leaf node of the decision tree to prescribe a duration for the phone being synthesized, Duong et al. used the mean and the standard deviation at the leaf node to compute the likelihood of the phone.

Figure 2 shows a fragment of the decision tree for phoneme duration, with the phoneme "IH" at the leaf node. Depending on the context that the phoneme appeared in, it is assigned a mean and standard deviation which is then used to score an utterance by using the actual duration of the phonemes in the utterance.

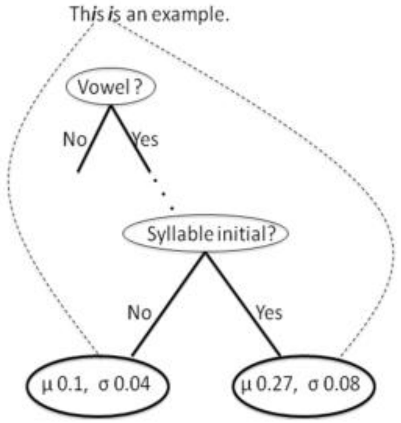

**Figure 2: Decision tree fragment with two /IH/ nodes**

The generalized model described in (Duong and Mostow, 2010) included sub lexical features like phoneme and syllable characteristics. The generalized model in (Duong, et al., 2011) only scored duration; we added features for pitch, pauses, power and punctuation to build pitch and intensity models in addition to the duration model. The duration model captured pauses implicitly, and all three models (duration, pitch and intensity) had punctuation features. We included only syntactic and lexical features, not sub-lexical features, so as to find differences in children's prosody in specific contexts interpretable in terms of reading rather than pronunciation, e.g. specific to types of words or syntactic constructs, not particular phonemes. Table 1 lists features used in the prosody models. In addition, we used the first two features of the previous and next words as lexical context features.

Table 1: Features used by narration models

| Feature class | Values |
|---|---|
| Part of speech of word | Penn Treebank POS tags (Pettibone) |
| Content word? | Binary |
| # of syllables in word | Number |
| Type of phrase | Noun, verb, etc. |
| Position of word in phrase | Number |
| Phrase break after word? | Binary |
| Punctuation after word? | All punctuation marks |

We built models of phoneme pitch, intensity and duration on a corpus of 4,558 sentences read by 11 fluent adults. For each phoneme in each child's utterance, we obtain the path taken by the phoneme through the three decision trees and the log likelihood of the phoneme duration, intensity or pitch being produced by the adult.

We scored each feature (or node in the path taken by the phoneme through the decision tree) by aggregating the log likelihoods obtained at the leaves and normalizing for the number of times each node was visited, by allocating credit to each node that was visited by the phone. This gave us a score for each feature for each child. Nodes in the tree that were not visited by any of the paths taken by the child's phonemes were given a value of 0. There were a total of 158 features for pitch, 115 features for intensity and 166 features for duration.

We also scored each feature across all children by aggregating the score for a particular feature for each child.

To reduce the noise in the data, we filtered out off-task sentences from our data set of 141,413 sentences read by children who used Project LISTEN's Reading Tutor (Mostow and Beck, 2007) in the 2005-06 school year. An off-task sentence is an utterance other than an attempt to read the sentence text (Chen and Mostow, 2011). After filtering, 85,209 sentences remained, read by 55 children, ranging from 54 to thousands per child.

## 4. Results

We now characterize how the children scored on prosody.

### 4.1 How did children's prosody most resemble adults'?

Table 2 lists the features that scored highest when we aggregated scores for each feature across all 55 children.

Table 2: Features that children scored best on

| Pitch | Intensity | Duration |
|---|---|---|
| Wh- pronoun, e.g. *who* | Past first 5 words (of phrase) | Before a present participle, e.g. *is jumping* |
| Cardinal number | Wh-pronoun | After a particle, e.g. *the boy* |
| Modal, e.g. *must* | Not followed by punctuation | First 4 words |
| 1st or 2nd word in phrase | Adjective, e.g. *beautiful* | Wh-pronoun |
| Adverb, e.g. *beautifully* | 1-2 syllable word, e.g *cooked* | 1-syllable word, e.g. *cook* |

As one might expect, children scored better on common monosyllabic and disyllabic words, such as particles, wh-pronouns, cardinal numbers, and present participles. Words early in a phrase scored higher.

### 4.2 Which features did children score worst on?

Table 3 lists the features that scored lowest when we aggregated scores for each feature across all 55 children. Some of these findings match Miller and Schwanenflugel (2006). They found smaller end-of-sentence pitch declination for children than for adults, which could explain why children did not score well on words that were followed by a period, and when there was a phrase break after the word. Some other findings make linguistic sense:

- Determiners are function words, normally unstressed.
- Fluent prosody elongates phrase and sentence endings.
- Long words tend to be informative and hence stressed.
- Things important enough to count often get emphasis.
- Pitch tends to be lower at unstressed function words.
- Most English phrases have alternating stress patterns.

Table 3: Features that children scored worst on

| **Pitch** | **Intensity** | **Duration** |
|---|---|---|
| End of sentence | Before singular or mass noun, e.g. *too many* | 1st or 2nd person singular present verb, e.g. *give* |
| Past tense verb, e.g. *divided* | After determiner, e.g. *the house* | Before phrase break, e.g. … *democracy, …* |
| Wh-adverb, e.g. *when* | 3rd person singular present verb, e.g. *gives* | After personal pronoun, e.g. *she gives* |
| After function word | 4+ syllable word, e.g. *caterpillar* | Cardinal number |
| After personal pronoun | 1st word in phrase | End of sentence |

### 4.3 Did prosody predict fluency and comprehension?

Stepwise linear regression using SPSS (SPSS, 2000) on children's aggregated scores for each feature selected 7-16 features in each of the final models to predict their fluency and comprehension test scores and gains. As Table 4 shows, models trained on pitch, intensity and duration and all three aspects of prosody predicted test scores much better than the template model (Mostow and Duong, 2009) and the generalized model that used only duration (Duong, et al., 2011). Our models likewise predict fluency scores more reliably than comprehension scores. However, we predict comprehension gains more reliably than fluency gains. Table 5 through Table 8 list the top 5 lexical and syntactic features selected by stepwise linear regression for predicting test scores and gains.

Table 4: Adjusted $R^2$ of all the models

| | **Pitch** | **Int** | **Dur** | **All** | **Tem-plate** | **Generalized** |
|---|---|---|---|---|---|---|
| **Fluency:** | | | | | | |
| **Scores** | .713 | .604 | .777 | **.810** | .565 | .572 |
| **Gains** | .587 | .532 | .502 | **.694** | - | - |
| **Comprehension:** | | | | | | |
| **Scores** | .599 | .503 | .711 | **.729** | .362 | .369 |
| **Gains** | .747 | .690 | .652 | **.787** | - | - |

Table 5: Features that best predict fluency scores

| **Pitch** | **Intensity** | **Duration** |
|---|---|---|
| First 2 words (of phrase) | Before present participle | First 5 words |
| Plural, e.g. *cats* | 1-2 syllable word | Content word |
| 1-2 syllable word | Before modal word | Coordinating conjunction e.g. *and* |
| After particle | After comparative, e.g. *better* | After plural noun |
| Before proper noun e.g. *says Michael* | Before plural noun | After 3rd person singular present verb eg. *bring it* |

Table 6: Features that best predict fluency gains

| **Pitch** | **Intensity** | **Duration** |
|---|---|---|
| First 3 words | After Wh-pronoun | Past participle |
| Before base form of verb, e.g. *I know* | After coordinating conjunction | First 4 words |
| After function word | 1-2 syllable word | Proper noun |
| Present participle | Before particle | First 2 words |
| In noun phrase | Proper noun | In verb phrase |

Table 7: Features that best predict comprehension scores

| **Pitch** | **Intensity** | **Duration** |
|---|---|---|
| Plural noun | Past participle verb e.g. *jumped* | Cardinal number |
| First 2 words | Before phrase break | Auxiliary verb, e.g. *"did"* |
| Modal verb | 1-2 syllable word | Before "," |
| 1 syllable word | Before symbol | First 5 words |
| Before past participle | First 3 words | Coordinating conjunction |

Table 8: Features that best predict comprehension gains

| **Pitch** | **Intensity** | **Duration** |
|---|---|---|
| First 4 words | 1-3 syllable word | After proper noun |
| 1-2 syllable word | Before particle | After coordinating conjunction |
| Before function word | Present participle | 1-2 syllable word |
| After plural noun | Before phrase break | Before cardinal number |
| Personal pronoun | After coordinating conjunction | No punctuation following word |

Inspection of which features appear in Table 5 through Table 8 suggests that prosody of short words (at most 2-3 syllables) and the first 2-3 words in a phrase predicts fluency and comprehension scores and gains better than prosody on longer and later words. This effect might indicate linguistic effects such as better prosody on the head of a phrase than on its subordinating clauses. On the other hand, it may only reflect a skewed sample of data on long words and phrases, with more proficient readers and hence lower variance. Comparing the features in Table 5 and Table 6 to the features in Table 7 and Table 8 suggests that prosody at punctuation and phrase breaks predicts comprehension better than fluency, perhaps because it indicates syntactic processing essential to comprehension.

## 5. Conclusion

We developed, implemented, and applied a method to mine 85,000 children's utterances scored against models of adult reading pitch, intensity and duration. Out of hundreds of text features, we identified the features on which children scored best and worst, and, more importantly, which features best predicted their fluency and comprehension. Besides predicting fluency and comprehension test scores dramatically better than previous published models, they (unlike our prior work) also predict pre- to post-test gains.

Such early indicators of progress are crucial in identifying students who need extra help (NCRI, 2010). The most predictive features might shed light on skill acquisition and might be useful pedagogical targets. More sensitive indicators of growth may help better analyze what kind of reading practice helps whom, and when (Beck and Mostow, 2008). Prosody is an audible indicator of comprehension. Thus mining oral reading prosody data may identify specific kinds of words and text that human and automatic tutors can listen for and teach in order to improve children's oral reading fluency more effectively.

One limitation of this work is that it is based on speech collected from assisted reading of text presented on a computer screen sentence by sentence, in order to give the Reading Tutor an opportunity to give feedback before going on. This mode does not capture the transitions between sentences characteristic of connected text – such as poor readers' failure to pause at sentence boundaries. Analyzing such transitions would require redesigning the Reading Tutor to let children read across sentence boundaries, while preserving its ability to interrupt them.

Another limitation of this work is that our empirical findings are hard to interpret: we trained models that predict lower or higher scores or gain, but not the reasons why. In particular, we scored oral reading prosody based only on its likelihood in an adult model – not on whether the pitch, intensity, or duration was above or below the mean adult value, only on how far it deviated. Future work should distinguish whether a child read a word higher or lower, louder or softer, and shorter or longer than a fluent reader would.

Moreover, our current models use only shallow (lexical and syntactic) features of text. Consequently, they cannot critique the appropriateness of prosody to meaning, such as contrastive stress based on context. Future work should identify additional text features relevant to oral reading prosody, especially higher level information such as semantic features, types of pauses, and the sentence types used by Miller and Schwanenflugel (2006).

Making sense of our empirical findings is challenging, and the extensions proposed above may improve their interpretability. One reason is to advance scientific understanding, for example by generating meaningful hypotheses driven by data but informed by psycholinguistic literature. Another reason is to enrich feedback on oral reading prosody (Sitaram et al., 2011) from scoring prosody graphically as good or bad to explaining why, and what to do about it.

## Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080628. 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

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*, 353-362. Montreal.

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

Black, A., Taylor, P., and Caley, R. 1996-1999. The Festival Speech Synthesis System.

Boersma, P. and Weenink, D. (2008). Praat: doing phonetics by computer (Version 5.0.33). Retrieved from http://www.praat.org

Chen, W. and Mostow, J. 2011. A Tale of Two Tasks: Detecting Children's Off-Task Speech in a Reading Tutor. In *Interspeech: Proceedings of the 12th Annual Conference of the International Speech Communication Association*, Florence, Italy.

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 *Online Proceedings of the Second ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, at http://www.eee.bham.ac.uk/SLaTE2009/. Wroxall Abbey Estate, Warwickshire, England.

Duong, M. and Mostow, J. 2010. Adapting a Duration Synthesis Model to Score Children's Oral Reading. In *Interspeech 2010*, 769-772. Makuhari, Japan.

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), 7*(4), 14:11-22.

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.

Kuhn, M. R., Schwanenflugel, P. J., and Meisinger, E. B. 2010. Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency. Invited review article. *Reading Research Quarterly*, *45*(2), 230–251.

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., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., . . . 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. (2007). When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. In B. Schneider & S.-K. McDonald (Eds.), *Scale-Up in Education* (Vol. 2, pp. 183--200). Lanham, MD: Rowman & Littlefield Publishers.

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

NCRI. 2010. Essential Components of RTI – A Closer Look at Response to Intervention. Washington, DC: National Center on Response to Intervention.

NRP. 2000. Teaching children to read: An evidence-based assesment of the scientific research literature on reading and its implications for reading instruction (pp. 00-4769). Washington, DC: National Institute of Child Health and Human Development.

Pettibone, J., from http://bulba.sdsu.edu/jeanette/thesis/PennTags.html

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., Kuhn, M. R., Morris, R. D., and Bradley, B. A. 2006. The Development of Fluent and Automatic Reading: Precursor to Learning from Text Retrieved November 8, 2006, from http://drdc.uchicago.edu/community/project.phtml?projectID=60

Sitaram, S., Mostow, J., Li, Y., Weinstein, A., Yen, D., and Valeri, J. 2011. What visual feedback should a reading tutor give children on their oral reading prosody? In *Online proceedings of the Third SLaTE: ISCA (International Speech Communication Association) Special Interest Group (SIG) Workshop on Speech and Language Technology in Education*, http://project.cgm.unive.it/events/SLaTE2011/programme.html. Venice, Italy.

SPSS. (2000). SPSS for Windows (Version 10.1.0). Chicago, IL: SPSS Inc.