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

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

What visual feedback should a reading tutor give children on their oral reading prosody?

Sunayana Sitaram, Jack Mostow, Yuanpeng Li, Anders Weinstein, David Yen, Joe Valeri

Project LISTEN, School of Computer Science,
Carnegie Mellon University

{ssitaram, mostow, yuanpengli, andersw}@cs.cmu.edu, aviden@gmail.com, jmv@andrew.cmu.edu

Abstract

An automated reading tutor that models and evaluates children's oral reading prosody should also be able to respond dynamically with feedback they like, understand, and benefit from. We describe visual feedback that Project LISTEN's Reading Tutor generates in realtime by mapping prosodic features of children's oral reading to dynamic graphical features of displayed text. We present results from preliminary usability studies of 20 children aged 7-10. We also describe an experiment to test whether such visual feedback elicits oral reading that more closely matches the prosodic contours of adult narrations. Effective feedback on prosody could help children become fluent, expressive readers.

Index Terms: prosody, visual feedback, intelligent tutoring systems, children, speech technology for education.

1. Introduction

Fluency is the ability to read not only quickly, easily, and accurately, but expressively – that is, with prosody (timing, intonation, and stress) appropriate to the text. Fluency rubrics for teachers to score oral reading include expressiveness [Allington, 1983; Pinnell et al., 1995; Rasinski, 1990].

Text lacks the prosodic cues that speech provides, so young readers must learn to make up for them, for example by attending to punctuation or syntax [Schreiber, 1987]. Expressive reading by a teacher or parent provides a model of appropriate prosody [Rasinski, 2003]. There is more than one appropriate way to read any given sentence, but resemblance to adult prosody predicts performance and gains in fluency and comprehension [Miller and Schwanenflugel, 2006; Miller and Schwanenflugel, 2008; Mostow and Duong, 2009; Young et al., 1996].

Effective tutoring includes not only modeling and practice, but feedback as well. Making expressive, appropriate prosody an explicit goal should improve comprehension by getting children to attend to the textual features and comprehension processes required to read expressively instead of, word, by, word. But getting children to share this goal may require giving them feedback on their performance.

Feedback on prosody is difficult because prosody is both fleeting and invisible. Fortunately, technology affords novel forms of feedback inspired in part by the work of artist Golan Levin (www.flong.com) on interactive voice-driven displays and their success in engaging children's interest. These displays translate prosodic properties of speech, such as pitch, amplitude, and duration, into intuitive visual analogues such as position, color, and size, enabling children to create displayed shapes simply by speaking.

We adapt such technology to give visual feedback on reader prosody by dynamically modifying the displayed text to reflect it. Some feedback simply mirrors the child's prosody, while normative feedback reflects assessment of the child's prosody. The prosody-matching task should foster prosodic awareness, and will serve as a motivational reward in itself if children find it enjoyable.

The rest of the paper is organized as follows. Section 2 relates this work to prior research. Section 3 describes our approach. Section 4 reports initial user studies with children. Section 5 describes ongoing work. Section 6 concludes.

2. Relation to prior work

Bonneau and Colotte [2011 (to be published)] have studied visual feedback on oral reading, but mostly on pronunciation rather than prosody. This feedback has taken various forms.

Talking heads offer a naturalistic feedback mechanism for pronunciation and prosody [Engwall and Balter, 2007]. Project LISTEN's Reading Tutor's interventions include playing a short video clip of a mouth pronouncing a phoneme, illustrated in Figure 1 of [Aist et al., 2002]. Interaction with a human-like virtual agent enables interventions that would be impossible in the physical world, such as cutaway views of the vocal tract, lips, and articulators [Cole et al., 1999; Cylwik et al., 2008].

The Fluency Pronunciation Trainer [Eskenazi and Hansma, 1998] for foreign language pronunciation training used a speech recognizer to compare the student's vowel durations to those of native speakers reading the same prompt. The system displayed a symbol below each word to indicate whether its vowel segment was too long, too short, or OK.

Vardanian [1964] used melodic curve visualization for second language learning and tested its impact on learners, with disappointing results, perhaps due to the quality of the visualizations. In the system described in [Bonneau and Colotte, 2011 (to be published)] the spectrogram and F0 curve of the user's utterance are shown, with arrows indicating whether the pitch of the target syllables should be lowered or raised. The color of each arrow maps to the difference in height between the user's F0 and native realizations of the syllable. The display also includes a curve representing the F0 contour and bars representing the syllable and vowel durations of the narration and those of the user.

The closest work we found to visual feedback for children on oral reading prosody was GRAFYC [Finlay], a series of games that promote prosody awareness among young children. However, the feedback is not on children's oral reading. The system displays a sentence, plays a recorded narration of it, and prompts the child to click on the location of a prosodic feature such as a stress or pause. Game characters react when the child responds correctly.

To explore possible games based on children's oral reading prosody, we conducted design studies of target-hitting games, implemented as separate programs driven by children's oral reading previously recorded by the Reading Tutor. Each game displayed the words of a sentence as a series of targets for the game character – a morsel of cheese for a mouse to eat, a balloon for a paper airplane to pop, or a fly for a frog to catch – at heights corresponding to their pitches in the adult narration. The player's oral reading prosody controlled the character's trajectory. Hitting a target made it vanish. This feat seemed too hard even for fluent adult readers, suggesting the need for more holistic feedback at a higher grain size.

3. Approach

Visual feedback on oral reading prosody should serve multiple purposes. It should engage children in oral reading by letting them use it to control the display, and by responding dynamically. It should create a visible trace of children's oral reading prosody in a form intuitively understandable to children and their teachers. It should also, when appropriate, apply the same visual effects to the Reading Tutor's recorded narrations to make visible the appropriate prosody they model. It should challenge the student to read a given sentence in such a way as to match its displayed prosodic contour as closely as possible.

In order to explore a space of designs for prosody visualization, we developed a program to input oral reading, compute prosodic features using the Sphinx [CMU, 2008] front end, aggregate frame-level features into word-level features, and display them according to an input specification of how to map each feature. To modify the display, we simply edit this mapping, represented as an SQLite query that specifies which graphical features to use (e.g. rectangles or text words) and which prosodic feature to map to each graphical feature.

Initially we mapped each word's prosodic features to graphical features of a static display. As Figure 1 shows, one such mapping displayed pitch as vertical position, pause duration as horizontal spacing, and word duration (normalized by number of letters) as font size.

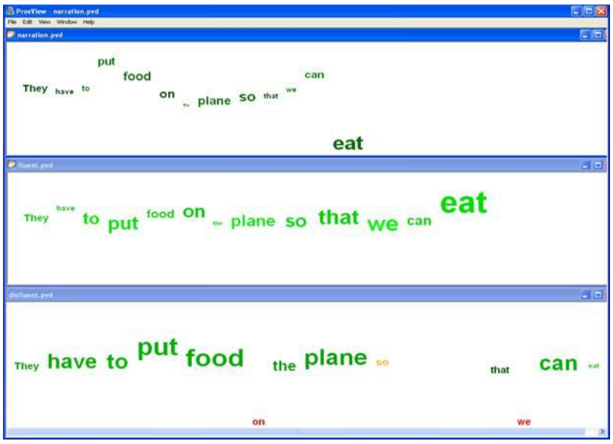
ProsView - narration.pvd
File Edit View Window Help
narration.pvd
They have to put food on the plane so that we can eat
fluent.pvd
They have to put food on the plane so that we can eat
disfluent.pvd
They have to put food the plane so that can eat
on we

Figure 1: Initial static display for adult narrator, fluent child, and disfluent child

We informally user-tested this mapping by displaying Figure 1 on a poster at an education conference and asking visitors to the poster to guess the mapping. They typically interpreted word height as pitch and space width as silence duration, but font size as intensity rather than word duration.

The next version still mapped the target adult narration to a static display, but displayed dynamic feedback on the child's prosody. Displaying both contours as words is confusing, so we mapped the narration to rectangles, providing a "cityscape" (adjoining rectangles of different heights) or "staircase" (horizontal lines at different heights) as a target pitch contour to aim at. To maximize transfer to normal reading, we laid out text as normally as possible. To emphasize the mapping from text words to prosodic contour, we preserved their horizontal locations, varying just their vertical positions, size, and color. We displayed prosodic contours below the current sentence, in the blank area where the next sentence will appear.

This design reflects several lessons learned along the way:

1. Don't map different features to word height and width. It distorts their aspect ratio, with ugly results.
2. Don't map word times directly to the x-axis. Slow reading with long pauses takes too much space. The x-axis may be the most natural way to display time in a static display. But in a dynamic display, the synchronicity of sound and graphics conveys time intuitively by revealing, moving, or highlighting each text word when it is read aloud.
3. Displaying sentences longer than one line is a problem; none of the following solutions is ideal:
   a. Squeezing a long sentence into a single line makes the text too small to read.
   b. Scrolling a long sentence horizontally makes it hard to read and even harder to reread.
   c. Splitting a contour across lines keeps text and contour together, but obscures the contour's overall shape and requires extra vertical space between lines to leave space for the contour. We picked this option as the least of the evils.

Table 1 describes the mapping from prosodic to graphical features for dynamic feedback in our preliminary user studies.

Prosodic feature of word	Graphical feature
Timing	Timing
Pitch	Height
Intensity	Font size
Latency	(Complement of) saturation
Confidence	Brightness

Table 1: Mapping prosody to dynamic graphical features

Pitch: As Figure 2 illustrates, we map a word's pitch to its vertical position. A word with no pitch estimate (e.g. omitted, rejected, or unvoiced) appears at the same vertical position as the previous word, in order not to look like a pitch excursion.

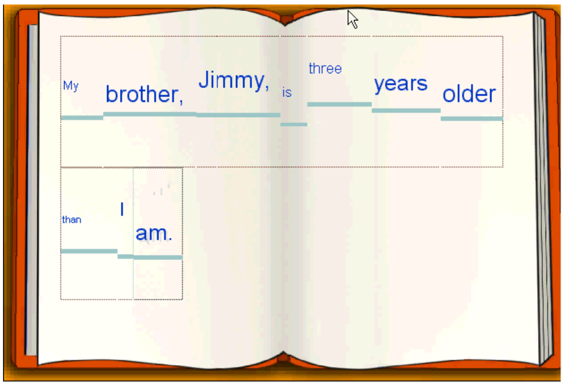

Figure 2: Words read with varying pitch and intensity

The Reading Tutor moves each word up or down according to how high or low the student spoke it, relative to his or her baseline average pitch. To adjust for differences in mean F0, we map the child's and narrator's baselines to the same vertical position. We estimate the narrator's baseline F0 by averaging over the sentence contour, which we know in advance. However, we do not know the child's mean F0 for the sentence in advance, so we use 270 Hz [Sorenson, 1989], which is typical for children. We plan to replace it with the F0 of the first word of the sentence or the mean F0 of the child's reading so far.

Intensity: As Figure 2 also illustrates, we map intensity to font size – the louder the word is read, the larger it appears.

Mapping a feature to font size can make words overlap, reducing legibility, so we constrain words to fit inside non-overlapping bounding boxes: a word can expand at most to where the next word begins.

Latency: To encourage fluent reading, we map larger latency to lower saturation. The longer the child pauses before a word, the paler it turns, e.g. the word with in Figure 3.

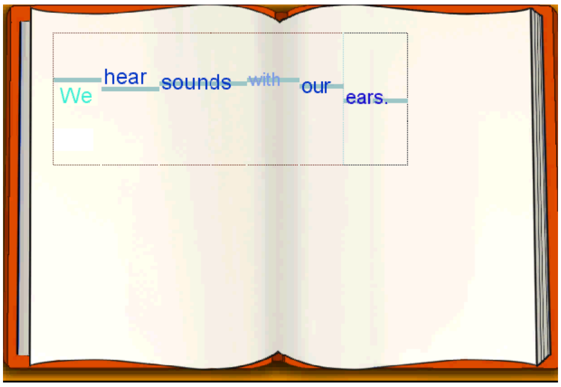

Figure 3: Words read with varying latency

Confidence: We map ASR confidence to brightness, so a low confidence or rejected word is darker or black.

Our preliminary user studies showed that children did not perceive the narration contour as a target. To emphasize this role, we now map target proximity to color. That is, the closer the child's pitch corresponds to the narrator's (adjusting for difference in mean pitch), the greener the word turns.

4. Evaluation

We now describe our evaluation procedure and results.

4.1. Preliminary usability testing

We conducted user studies on 20 children ages 7-10, in grades 2, 3 and 4. All students except one had used the Reading Tutor before. We did not tell the students what the prosody display meant, or how they were supposed to interpret the visual feedback. The preliminary user test aimed to find out:

1. How much information should we display?
2. What are children's mental models of the display?
3. What do they have to say about the feedback?
4. How would they like it to look and behave?

Table 2 lists the five variants we tested. Variants 1 and 2 compared two contour shapes to see if they preferred either shape or were likelier to perceive it as a target. Variants 2 through 5 compared successively larger amounts of information displayed.

Variant	Prosodic features mapped	Narration contour
1	Pitch, intensity	cityscape
2	Pitch, intensity	staircase
3	Pitch, intensity, latency	staircase
4	Pitch, intensity, ASR confidence	staircase
5	Pitch, intensity, latency, ASR confidence	staircase

Table 2: Variants used for usability testing

During the usability test, each student first read one or more stories. When a student finished a story, the Reading Tutor asked 4 multiple-choice questions.

The survey questions at the end of the stories asked the students whether they enjoyed using the new version of the Reading Tutor or not, whether it was easier or more difficult to read with the prosody display. Students were also asked if they understood why the words looked the way they did and whether they understood what the shapes under the words meant.

Afterwards, one of the authors interviewed the student for 2-3 minutes about his or her experience with the new display. During the interviews, we asked the children to describe what they saw on the screen and guess what the words moving up and down, the size of the words and their colors meant. We also asked them what they liked and didn't like about the way the words looked, and what colors or shapes they would use to indicate "good" or "bad" reading. We were careful in the interviews not to describe the actual mappings.

4.2. Observations and feedback

- 16 of the 20 children said in the interview that they liked reading with the prosody display more than reading without it, typically because it was "fun". 2 children said it was "ok," and 2 children said they didn't like it.
- Only 2 children realized that the rectangles and staircase at the bottom correspond to the way the tutor reads the sentence.
- Four children said (incorrectly) that the words moved up and down according to how fast they read, or that the words went up and down to show that they read the word correctly. That is, they thought that latency or accuracy mapped to vertical position, rather than pitch.
- They suggested using color to indicate good and bad reading. 8 children suggested using their favorite color for good reading. There wasn't a general consensus about which color could indicate good reading and which one could indicate bad reading, although some seemed to agree that green and red were good choices, respectively.
- Only 2 of the first 17 children could tell that intensity mapped to font size, but font size didn't vary enough to be very noticeable, so we adjusted the mapping to amplify the differences. 2 of the 3 children tested subsequently noticed the mapping. 3 of the earlier children thought that the word got bigger if they read it correctly and smaller if they got it wrong, i.e., that correctness mapped to font size.
- The children noticed that some words were paler or darker (due to latency and confidence) but did not know why. 3 children said that the colors changed when they got the word wrong.
- 2 children felt that words moving around made it harder for them to read and preferred the older version. One of the other children said that the words moving up and down made it easier for him to read the next word and not skip it.
- Errors by the Reading Tutor in tracking the child's position are more obvious when the words move. One child said that when multiple words moved down quickly because of a tracking error, he got confused about which word he was on. Some children skipped words that moved down due to being hallucinated by the ASR.
- When we asked some children what they wanted the display to look like when they read something wrong, 2 of them said that they wanted the Reading Tutor to correct them verbally but they did not want to see any change in display. We think the reason is that they did not want evidence of errors on the screen for others to see.

5. Work in progress

Section 4 focused on children's mental models, and whether they liked using the Reading Tutor with the prosody display. An experiment now underway tests whether visual feedback actually increases the resemblance of children's prosodic contours to adult narrations of the same sentences.

The experiment manipulates three independent variables:

1. Does the tutor play the adult narration?
2. Does the tutor display the static narration contour?
3. Does the tutor display dynamic feedback on prosody?

The control condition includes narration but not display. We will test static contours and dynamic feedback separately to assess whether dynamic feedback helps or hampers reading.

6. Conclusions

The paper sets the goal of engaging, understandable, effective graphical feedback on children's oral reading prosody. We frame the problem of mapping word-level prosodic features to graphical features, and discuss our explorations in the space and the design decisions based on them. We describe an implemented tool to generate feedback in the Reading Tutor. We present report observations and feedback from 20 children in preliminary usability testing.

Although it would be rash to draw firm conclusions from small pilot studies, they suggest some tentative hypotheses for future work to test: Children enjoy dynamic visual feedback. They generally don't understand the static target or dynamic mapping. Most children enjoy the moving words, but a few find them hard to read. Children enjoy color, especially as normative feedback. Even without seeing or understanding the static target, children seem both eager and able to approach it based on proximity color. Dynamic prosodic feedback seems to elicit prosody that resembles a fluent adult more than it would otherwise. We need more user testing to find the best combination(s) of the independent variables: whether to display the target adult contour (statically), the child's prosodic contour (dynamically), and/or its proximity to the target.

An important question for future research is when to introduce prosody feedback. Visual feedback may be distracting for a child who is struggling to decode words. It may make sense to delay prosody feedback until a child is comfortable with reading stories of a certain level.

Data collection has now commenced for our experiment to test the effectiveness of the prosody display in terms of helping children emulate adult oral reading prosody. If successful, such visual feedback could help improve oral reading fluency, expressiveness, and comprehension.

7. 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.

8. References

AIST, G., KORT, B., REILLY, R., MOSTOW, J. and PICARD, R. 2002. Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens Increases Student Persistence [Poster]. In *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems (ITS'2002)*, Biarritz, France, June 5-7, 2002, S.A. CERRI, G. GOUARDÈRES and F. PARAGUAÇU, Eds. Springer, 992.

ALLINGTON, R.L. 1983. Fluency: The neglected reading goal. *Reading Teacher 36*(6), 556-561.

BONNEAU, A. and COLOTTE, V. 2011 (to be published). Automatic Feedback for L2 Prosody Learning. In *Speech Technologies*, Intech.

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

COLE, R., MASSARO, D.W., VILLIERS, J.D., RUNDLE, B., SHOBAKI, K., WOUTERS, J., COHEN, M., BESKOW, J., STONE, P., CONNORS, P., TARACHOW, A. and SOLCHER, D. 1999. New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. In *ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*, London, UK, April, 1999.

CYLWIK, N., DEMENKO, G., JOKISCH, O., JÄCKEL, R., RUSKO, M., HOFFMANN, R., RONZHIN, A., HIRSCHFELD, D., KOLOSKA, U. and HANISCH, L. 2008. The use of CALL in acquiring foreign language pronunciation and prosody – General specifications for Euronounce Project. In *Proc. SASR*, Piechowice, Poland, September 2008, 2008.

ENGWALL, O. and BALTER, O. 2007. Pronunciation Feedback from Real and Virtual Language Teachers. *Computer Assisted Language Learning 20*(3), 235-262.

ESKENAZI, M. and HANSMA, S. 1998. The Fluency Pronunciation Trainer. In *Proc. Speech Technology in Language Learning*, Marholmen, Sweden, 1998.

FINLAY, S.A. GRAFYC: Game For Realizing Accentuation For Young Children.

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 DUONG, M. 2009. Automated Assesment of Oral Reading Prosody. In *14th International Conference on Artificial Intelligence in Education (AIED2009)*, Brighton, UK, July 6-10, 2009, 2009, V. DIMITROVA, R. MIZOGUCHI, B.D. BOULAY and A. GRAESSER, Eds. IOS Press, 189-196.

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.

RASINSKI, T.V. 1990. Investigating Measures of Reading Fluency. *Educational Research Quarterly 14*(3), 37-44.

RASINSKI, T.V. 2003. *The fluent reader: oral reading strategies for building word recognition, fluency, and comprehension*. Scholastic Professional Books, New York, N.Y. 192 pp.

SCHREIBER, P.A. 1987. Prosody and structure in children's syntactic processing. In *Comprehending oral and written language*, R. HOROWITZ and S.J. SAMUELS, Eds. Academic Press, San Diego, CA, 243-270.

SORENSON, D.N. 1989. A fundamental frequency investigation of children ages 6-10 years old. *Journal of Communicative Disorders 22*, 115–123.

VARDANIAN, R.M. 1964. Teaching English Intonation Through Oscilloscope Displays. *Language Learning 14*(3-4), 109-117.

YOUNG, A.R., BOWERS, P.G. and MACKINNON, G.E. 1996. Effects of prosodic modeling and repeated reading on poor readers' fluency and comprehension. *Applied Psycholinguistics 17*, 59-84.