The Sounds of Silence: Towards Automated Evaluation of Student Learning in a Reading Tutor that Listens

Jack Mostow and Gregory Aist

Project LISTEN, Carnegie Mellon University
215 Cyert Hall, 4910 Forbes Avenue
Pittsburgh, PA 15213
http://www.cs.cmu.edu/~listen
mostow@cs.cmu.edu, aist+@andrew.cmu.edu

Abstract[1]

We propose a paradigm for ***ecologically valid, authentic, unobtrusive, automatic, data-rich, fast, robust,*** and ***sensitive*** evaluation of computer-assisted student performance. We instantiate this paradigm in the context of a Reading Tutor that listens to children read aloud, and helps them. We introduce ***inter-word latency*** as a simple prosodic measure of assisted reading performance. Finally, to validate the measure and analyze performance improvement, we report initial experimental results from the first extended in-school deployment of the Reading Tutor.

Content areas: computer aided education, spoken language understanding, user interfaces, cognitive modeling, multimedia

Introduction

Tutors that listen can exploit a much wider and more natural input channel from their students than the keyboard and mouse used in most educational software. They have the potential to realize a novel paradigm for analyzing student performance, in which evaluation is:

- ***ecologically valid*** – done “in vivo” in the student’s normal school or home setting, not just in a lab
- ***authentic*** – applied to normal activities relevant to student goals and interests, not just artificial test materials
- ***unobtrusive*** – conducted “non-invasively” while the student is performing with assistance by the tutor, in contrast to conventional probes and pre- and post-tests
- ***automatic*** – without human supervision, assistance, intervention, transcription, scoring, or interpretation
- ***fast*** – computed in real-time on a PC
- ***data-rich*** – based on a large data sample
- ***robust*** – tolerant of inaccuracies in speech recognition
- ***sensitive*** – able to detect subtle, fine-grained, aggregated, or longitudinal effects

We instantiate this paradigm in a context where listening is especially important – tutoring children’s oral reading.

Importance of Oral Reading

At present, children are taught to read aloud in grades 1-3, and are expected to read silently by grade 4. Children who fail to read independently by grade 4 tend to fall further and further behind their classmates as they grow older, and are at substantial risk of growing up illiterate.

Oral reading is taught by a combination of classroom instruction and individual practice. Reading aloud helps children learn to identify printed words by relating them to the spoken form they have already learned. At this stage, children's comprehension of spoken language is typically years above their independent reading level (Curtis 1980).

Listening to children's oral reading is important for several reasons. First, it can detect word identification errors, so they can be corrected. McCoy and Pany (1986) summarized work on corrective feedback during oral reading: correcting word reading errors enhances word recognition accuracy and comprehension for students with learning disabilities. Moreover, studies of spoken assistance on demand (McConkie & Zola 1987, Reitsma 1988, Olson & Wise 1992, Lundberg & Olofsson 1993) have revealed a serious flaw in assuming young readers are willing and able to ask for help when they need it. Children with reading difficulties often fail to realize when they misidentify a word. Second, listening can detect disfluency – slow, halting reading likely to be associated with growing frustration and/or failing comprehension. Third, the very act of listening can have a powerful motivational effect, by giving young readers a supportive audience for their attempts at oral reading. Fourth, listening can be used to detect successes, not just mistakes – to identify what the child knows, and to provide positive reinforcement when the child succeeds.

Previous Work

Advances in technology have made the application of speech recognition to oral reading increasingly feasible, but its effectiveness has been evaluated manually.

Detection of reading mistakes and mispronunciations has been evaluated against human transcripts (Bernstein et al. 1990, Phillips et al. 1992, Mostow et al. 1994, Russell et al. 1996). Although useful for evaluating recognition accuracy, this method suffers from the expense of human transcription, the sparseness of mistakes compared to correct reading, the subjectivity of judgments about what constitutes a mistake, and the irrelevance of most mistakes to successful comprehension. Moreover, these evaluations show this automated detection compared poorly to human listening, with many false alarms and undetected mistakes.

Educational effectiveness has been evaluated by comparing student performance before and after tutor use (Kantrov 1991) or with and without the assistance of a tutor (Mostow et al. 1994). For example, one experiment measured how well 34 second graders comprehended a third-grade story they read with the assistance of an automated coach, compared to a similar story they read without assistance. Unassisted and assisted conditions counterbalanced order and story.

This evaluation illustrated several difficulties. Data collection required weeks of work to train the subjects, videotape use of the coach, administer the comprehension tests, and grade the answers. Comprehension scores for a story, based on only eight questions, were vulnerable to noise. Inter-student variability was too great to expect significant differences in average comprehension between the two conditions. To reduce within-subject variability, individual effect size for a subject was defined as:

(assisted comprehension score) - (unassisted comprehension score)

The mean effect size was positive, but whether it was statistically significant depended on which grader scored the answers, due to inter-grader differences caused by the subjective nature of the grading task. Finally, although the two stories were rated at the same difficulty level in the reading test they were taken from (Spache 1981), unassisted comprehension turned out to be considerably lower on one than on the other. Thus assisted comprehension was about 40% higher than unassisted comprehension for the “harder” story, with no significant difference for the “easier” story.

A separate evaluation sampled the videotape to identify usability problems. Problems were analyzed by inspecting detailed event logs of the coach’s responses, identifying the input (mouse clicks and speech recognizer outputs) that triggered it, and hand-transcribing speech input as needed. Three main sources of problems were identified. Errors by the speech recognizer could cause the coach to reject correctly read words, accept incorrectly read words, supply the wrong word when the reader got stuck, or go on to the next sentence prematurely. Delays of a few seconds between read word and coach response, besides slowing down overall progress through the text, sometimes caused such behavior as belatedly supplying a word after the reader had already recovered from getting stuck on it. Ambiguities in the interface occasionally confused readers as to whether they were expected to read a single word or the rest of the sentence. Mostow et al. (1995) concluded that the user should have more control, to accommodate a wider range of reading ability and individual differences.

Lab experiments in psychology have measured and analyzed reaction times for many tasks, including reading isolated words (e.g., Plaut et al. 1996). However, use of timing information to assist reading has been rare in the absence of speech recognition. An interesting exception was a system (L’Allier 1980) that “enabled poor high school readers to comprehend texts “as well as good readers reading the same texts on printed paper without assistance” (Reinking & Bridwell-Bowles 1991) by dynamically adjusting the level of reading material based on measures including reading time and response time for interspersed questions.

A Reading Tutor that Listens

We are developing an automated Reading Tutor using the speech analysis methods in (Mostow et al. 1994) and the design recommendations in (Mostow et al. 1995). Unlike the reading coach in (Mostow et al. 1994), which required a NeXT machine for the user and a Unix workstation for the speech recognizer, the Reading Tutor runs in Windows™ 95 or NT 4.0 on a Pentium™, with a Knowles noise-cancelling headset microphone. This platform is cheap enough to put in a school long enough to help children learn to read better. The Tutor incorporates materials adapted from *Weekly Reader* and other sources.

In October, 1996, we deployed the then-implemented portion of the Reading Tutor in an inner-city elementary school for a pilot study of extended use. The purpose of this pilot study was to explore the Tutor’s usability and effectiveness, identify opportunities for improvement, and generally see what would happen. At this point the Reading Tutor already incorporated months of design iteration based on usability tests, conducted in our lab and during a 2-day visit in June 1996 to the school, as well as lessons gained from user tests by hundreds of children of its predecessor the reading coach. However, in these tests children used the Tutor less than an hour, and only in the presence of the researchers. Would the Tutor be usable and robust enough for students and school personnel to operate it on their own? Would children be willing and able to continue using the Reading Tutor regularly over a period of weeks or months? Would their reading improve?

At the school's suggestion, the eight pilot subjects selected were the two lowest readers in each of its four third grade classrooms. These children read 1-2 years below grade level and were at greatest risk of growing up illiterate. The pre-test for the pilot study consisted of national standardized tests given each fall, plus detailed individual reading assessments performed by the school's reading specialist. A post-test is planned for spring 1997.

Each pilot subject was scheduled to use the Reading Tutor 30 minutes a day, modulo school schedule conflicts, student absences, and program crashes. Children were taken from class one at a time to use the Reading Tutor in a small room, escorted and supervised by a school aide. The Reading Tutor recorded detailed event logs and speech (minus silences) on removable 1GB Jaz™ disks.

Teacher feedback and student experience with the initially deployed version led to a few changes. Some were technically trivial but educationally important, such as changing "gr1, gr2, gr3" to "A, B, C" in the story names to avoid the stigma of reading below-grade-level material. Other changes were harder, such as reducing the impact of false alarms by modifying the algorithm for deciding when to go on to the next sentence (Aist, 1997).

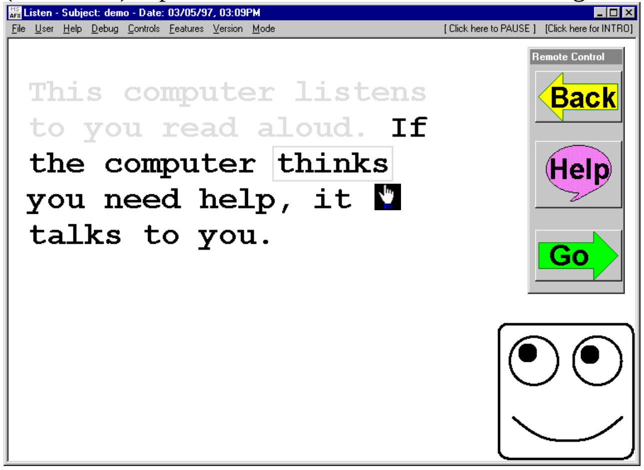
The resulting 11/7/96 version, running on a 90MHz, 64MB Pentium under Windows 95, was used throughout the pilot study reported here, in order to avoid variability or novelty effects due to version changes. This version suffered from some recognizer lag, but its simpler design appeared to avoid its predecessor's delay-related problems. (Subsequent versions have improved speed, accuracy, and functionality, for example by reducing memory usage to speed up the recognizer, converting to full-duplex to reduce truncation errors, and expanding the repertoire of user and Tutor actions.)

Design of the Reading Tutor (11/7/96 Version)

The Reading Tutor listens to a child read one sentence at a time. The Tutor displays an animated persona that actively watches and patiently listens, as shown in the screen shot. When the Tutor hears the end of the sentence or a prolonged silence, it aligns the speech recognizer output against the sentence to decide if any important words were missed. If not, it goes on to display the next sentence. Otherwise, the Tutor responds expressively by using recorded human voices. Responses include:

- **Supply** a word by saying "This word is <word>."
- **Recue** a word by speaking the text that precedes it.
- **Read** the sentence if more than one word was missed.

Then the Tutor lets the child reread the word or sentence. (Aist 1997) reports how the Tutor decides when to go on.


Listen - Subject: demo - Date: 03/05/97, 03:09PM
File User Help Debug Controls Features Version Mode
[Click here to PAUSE] [Click here for INTRO]
Remote Control
Back
Help
Go

This computer listens to you read aloud. If the computer thinks you need help, it talks to you.

In the real example shown here, the first sentence has just been recognized as read correctly, so the Tutor has grayed it out and displayed the second sentence. The child misreads it as *"if the computer...takes your name...help it...take...s to you."* The recognizer hears "IF THE COMPUTER THINKS YOU IF THE HELP IT TO TO YOU." The Tutor detects multiple missed words ("need" and "talks"), so it reads the sentence aloud and lets the child try it again. A box around "thinks" shows which word the cursor is on; the Tutor will speak this word if the child clicks. The Tutor provides additional learner control with buttons for:

- **Back**: Return to the previous sentence.
- **Help**: Speak the current sentence.
- **Go**: Go on to the next sentence.

This design accommodates individual differences better than the coach, by giving the student more control over:

What to read: Back and Go let the reader navigate. Perfectionists can read a sentence until the Tutor accepts it; impatient readers can read what happens next.

Who reads what: Novice readers can let the Tutor read a sentence first; bolder readers can try it first on their own.

How much to hear: The child can click to interrupt.

How to respond to correction: The Reading Tutor lets the child reread a corrected word by itself or in context.

What are typical sessions with the Reading Tutor like? From the sessions we have observed, children often attempt to read each sentence, but may click Help to have the Tutor read it first. They often click on words to hear the Tutor speak them, but misread many other words or get stuck in mid-sentence. The Tutor's typical correction is to speak the sentence. Children generally echo the sentence, which the Tutor may or may not then accept. Thus disfluent readers often wind up attempting the sentence twice or more, thanks to a combination of reading mistakes, speech recognizer errors, and a surprising tendency for children to repeat the sentence until the Tutor accepts it, rather than clicking Go.

The cognitive value of sentence repetition is unknown. Levy et al. (1993) found benefits of repeated reading, but their unit of repetition was a multi-sentence passage too long to hold in short-term memory. Sentence echoing probably helps comprehension (Curtis 1980) by reducing the cognitive load of identifying the words. Whether such repetition helps word identification may depend on whether children are just echoing the Tutor or attending to the mapping between the printed and spoken words.

The causes and motivational aspects of repetition are intriguing. Perhaps children don't understand or remember the Go button, but we suspect they perceive the interaction as a challenging game (Lepper et al., 1993) in which the goal is to get the Tutor to accept their reading. Despite occasional frustration, children seem to enjoy using the Reading Tutor, immediately become totally engaged, stay on task, and want to return to use it again.

A Novel Measure of Reading Performance

To evaluate and improve the Reading Tutor's effectiveness at helping children learn to read better, we will need to measure improvements in their performance. Ideally the Tutor should be able to measure such improvements itself, using just the data it normally captures. Why? Appropriate tutor feedback has been shown to increase performance and decrease time required for learning for a number of domains (Anderson et al. 1995). In order to provide appropriate assistance, a tutor must be able to assess a student's performance. According to Martin & VanLehn (1996), "A cognitive assessment system should: (a) integrate data from multiple activities, (b) analyze the data in a statistically sound, defensible manner, and (c) provide assessments at multiple grain sizes."

We have already discussed limitations of measures used to evaluate previous applications of speech recognition to oral reading. What might work better?

One candidate is ***reading rate***, defined as the number of correctly read words per minute. For the population of interest, this measure has a striking 0.9 correlation with reading comprehension (Deno 1985). Reading rate can be estimated accurately from a brief sample of oral reading, and is more sensitive than conventional comprehension tests to small improvements in reading ability (O'Connor, personal communication). However, reading rate varies with the level of difficulty of the material being read, and is too coarse-grained to measure individual word learning.

Conventional ***word latency*** measures the difficulty of reading a given word (in isolation) by the time interval from when the word is presented to when the reader starts to speak it. Laboratory studies have shown that this measure reflects word frequency and irregularity (Plaut et al. 1996). For our task we extend this measure to handle:

Connected text: Measure the time interval from the end of the previous word *i*-1 to the start of text word *i*. This measure is defined only when both words are read, so it is undefined for word 1 of the sentence, or for missed words. Start and end times for text words are computed by using a dynamic programming algorithm to align the text against the time-labelled output of the speech recognizer.

Disfluency: Include false starts, sounding out, repetitions, and other insertions, whether spoken or silent. Thus voiced attempts are not favored over silent rehearsal.

Assistance: Include time spent requesting help (deciding to ask, moving the cursor, and clicking) but not time during which the Tutor is responding. Thus only the reader's behavior is measured, not the help that affects it.

Inter-word latency averaged over a text is closely related to reading rate, since summing the latencies of all the text words yields the total length of the utterance, omitting sentence-initial and -final silences, unaligned words, and the time to speak each word. These omissions reduce variability due to machine-dependent Reading Tutor response time, word length, and speech production.

Text	IWL	Spoken	Recognized	Start	End
			<silence>	1	35
They	--	*they*	THEY	36	128
			<silence>	129	131
have	4	*have*	HAVE	132	188
			<silence>	189	207
learned	20	*learned*	LEARNED	208	281
			<silence>	282	309
something	29	*something*	SOMETHING	310	390
		a- ab-	THE	391	472
			<silence>	473	615
new	226	*new*	NEW	616	690
			<silence>	691	722
about	33	*about*	ABOUT	723	793
			<silence>	794	834
the	42	*the*	THE	835	878
			<silence>	879	927
		stregs-	START_ STEGOSAURUS	928	999
			<silence>	1000	1011
		sssss-	START_ STEGOSAURUS	1012	1039
			<silence>	1040	1100
		stegosaurus	STEGOSAURUS	1101	1231
			<silence>	1232	1315
Stegosaurus.	438	*stego?*	STEGOSAURUS	1316	1394
			<silence>	1393	1436

To clarify, this real example shows a sentence, the inter-word latency computed for each word, what the child said, and what the Tutor recognized. Times are in centiseconds. Long latencies reflect struggles at the words "new" and "Stegosaurus." Two apparent false starts on "about" are misrecognized as THE without affecting computation of the 226 centisecond latency before "new." The reader's struggles with the word "Stegosaurus" include two false starts, a correct attempt, and a final uncertain attempt misrecognized as STEGOSAURUS and aligned against the text word, yielding a somewhat inflated latency of 438 centiseconds. Nonetheless, this value reflects the reader's difficulty – better, in fact, than the 60 centisecond silence that immediately precedes the correct attempt.

Individual latency values are too variable to rely on – not just because of imperfect speech recognition, but because a long repetition (such as restarting the sentence) yields an anomalously large latency (even tens of seconds) for the ensuing word. To alleviate variability, we aggregate over multiple observations.

Evaluation

We performed a series of experiments to validate the inter-word latency measure, evaluate students' improvement, and analyze where improvement occurred.

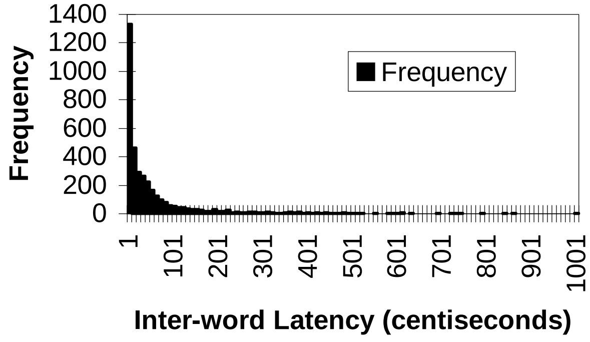
Experiment 1: Is the measure valid?

Before evaluating student learning or Reading Tutor effectiveness, we had to validate how well inter-word latency can measure difficulties in reading. It might fail in various ways. For example, applying word latency to connected text assumes that readers tend to decode and speak words in sequence. This assumption fails if, say, readers silently decode an entire sentence before speaking it. Or speech recognition might yield results too noisy to be useful. Or the Tutor's assistance might distort latency.

Expert validation compares inter-word latency against human judgment of where an individual reader has difficulties. Such comparison benefits from human listening and common sense, but is subjective, expensive, and imprecise. We did conduct an informal spot-check on a small sample, which showed agreement in most cases. Next we validated the measure objectively, economically, and quantitatively on a much larger sample.

Data. This test was based on sessions recorded Nov. 8-20, 1996, without the researchers present. This period included nine school days, with 3-7 sessions for each pilot subject. Each subject read from 32 to 212 text sentences totalling 160 to 1780 words, producing 71 to 379 utterances, including initial attempts, rereadings, and other responses to corrective feedback. The number of utterances per sentence ranges from 1.6 to 2.3, which includes the reader's initial attempt plus responses to subsequent Tutor correction(s). This range indicates feedback on a high proportion of sentences.

To control for Tutor assistance, we consider only a subject's first attempt at each sentence, omitting subsequent attempts (656 utterances) in order to exclude "echoes." The 794 remaining utterances correspond to 5956 presented words of text, 4824 (81.0%) of them aligned against recognizer output. 793 aligned words are sentence-initial, and 155 are preceded by an unaligned word, leaving a total of 3876 defined inter-word latencies. Their distribution looks exponential, with a large spike at 1 centisecond (the limit of resolution) and a long, thin tail:



Hypothesis. To validate the measure, we tested a prediction based on knowledge about reading – namely, that inter-word latency can distinguish known "easy" from "hard" words.

Procedure. We compared latencies for stopwords vs. nonstop-words. Mostow et al.'s (1994) 36 stopwords (a, all, an, and, are, as, at, be, by, for, he, her, him, his, I, if, in, is, it, its, me, not, of, off, on, or, she, so, the, them, then, they, this, to, we, you) cover about 50% of Spache's (1981) text.

Results. Mean inter-word latency was 70.2 centiseconds for the 1094 occurrences of stopwords, versus 94.7 centiseconds for the 2782 words not on the stopword list. A two-tailed t-test, assuming unequal variances, showed a significant difference between stopwords and non-stopwords at a 99% significance level (p-value = 0.0046).

Discussion. The inter-word latency measure is able to distinguish between "easy" and "hard" words, so we can use it to help detect improvement in student performance.

Experiment 2: Does performance improve?

Does performance improve with Tutor use? We could measure changes in reading rate, which would detect transfer of improvement in reading skill to new text. However, independent reading differs as a task from Tutor-assisted reading, where overall reading rate may be slowed down by Tutor assistance and sentence rereading. Also, reading rate is affected by the difficulty of the text, and some subjects chose harder stories as they progressed. Instead, we tested for improvement within-subject and within-word using our inter-word latency measure.

Data. This test was based on sessions recorded 10/22/96 - 3/21/97. Subjects had from 14 to 28 sessions. Absences, vacations, and downtimes up to two weeks limited usage. Excluding transit and startup overhead, sessions averaged 18 minutes of time on task from first to last utterance. Classroom use should reduce overhead and increase time on task.

To control for Tutor assistance, we again consider only a subject's first attempt at each sentence. The remaining utterances correspond to 23575 presented words of text, 20091 (85.2%) of them aligned against recognizer output, of which 14041 have defined inter-word latencies. To control for learning to operate the Tutor, we excluded data prior to 11/8/96, except to determine which sentences had already been presented.

Hypothesis. Inter-word latency decreases with Tutor use.

Procedure. To control for subject variability, we compared performance within rather than across subjects. To control for word variability, we calculated ***change in latency*** from a subject's first to last encounter of a word.

Results. Mean latency change (-37.5 ± 13.8) is significant at 95% overall. Improvement was significant at 95% for 3 subjects, at 90% for 3 subjects, and insignificant for one subject, with one subject insignificantly worse.

Discussion. Students are improving, but where?

Experiment 3: Does improvement transfer?

To characterize student learning, we analyzed where improvement occurred.

Hypothesis. Learned words transfer to new contexts:

Locally: speedup occurs over encounters close in time.

Persistently: improvements last over time.

Procedure. To control for story memorization, we looked only at word occurrences in new contexts. To reduce effects of prior learning, we excluded stopwords.

To test for local transfer, we compared students' first and last encounter of a word (in new contexts) on a given day. To test for persistent transfer, we compared students' first-ever encounter of a word (in the Tutor) against their first encounter on the last day they saw the word in a new context.

Results. Overall difference between first encounters and other encounters on the same day (-30.3 ± 24.9, n=137) is significant at 90%. Improvement was significant at 95% for one subject and at 90% for 2 subjects, with no significant change for the other. Overall difference between the first-ever encounter and first encounter on the last day (4.8 ± 30.5, n=259) was not significant.

Discussion. These results support the local transfer hypothesis, with 32.7% relative improvement in latency when students encountered recently seen words in new contexts.

Inter-word latency is sensitive enough to detect promising improvement on recent words. Why did we not find persistent learning? In most cases students had encountered the words in question in just a few contexts; perhaps students need more Tutor sessions or more varied exposure to a word for persistent improvement. Our measure has high variance – maybe we need more data or a refined measure. Finally, the deployed version of the Tutor lacked explicit, individualized instruction in domain skills, which has been found to lead to persistent learning in other domains (Anderson et al. 1995).

Conclusion

What does this paper contribute? The Reading Tutor extends its predecessors to support the extended use required for learning, as evidenced by months of use by a pilot group of poor readers, and measurable improvement.

The proposed paradigm addresses some limitations of previous methods to evaluate student learning. We showed how automatic speech recognition makes it possible to instantiate this paradigm in the context of the Reading Tutor. We introduced inter-word latency as a simple but useful prosodic abstraction to measure word difficulty. We measured changes in latency for the same word to help tease apart effects of student progress, text difficulty, word recency, and story memorization.

Does such evaluation live up to our stated criteria? Is it:

- ***ecologically valid*?** Reading is recorded at school, with normal supervision but no researchers required.
- ***unobtrusive*?** Assisted reading is evaluated invisibly.
- ***authentic*?** Text is student-selected, not a special test.
- ***automatic*?** Just align text against recognizer output.
- ***fast*?** Latencies are computed in real-time on a PC.
- ***data-rich*?** Word latencies are captured at the student's reading rate, giving O(10 data points/min.).
- ***robust*?** Latency tolerates many recognizer errors.
- ***sensitive*?** Latency is too noisy to rely on individual values, but detects subtle effects on subsets of words.

The Reading Tutor is working in a school setting, and children are using it. We can now study not just how well it helps children read, as earlier systems did, but how well they learn over time. Automated measures like the one reported here should help answer future key questions: How well are children learning? Is the Tutor helping them learn? How can it help better?

Acknowledgments. This material is based upon work supported in part by the National Science Foundation under Grant No. IRI-9505156 and CDA-9616546, by the Defense Advanced Research Projects Agency under Grant Nos. F33615-93-1-1330 and N00014-93-1-2005, and by the second author's National Science Foundation Graduate Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government.

We thank Drs. Leslie Thyberg, Rollanda O'Connor, and Gayle Griffin for expertise on reading; Raj Reddy and the CMU Speech Group (especially Ravi Mosur) for Sphinx-II; Alex Hauptmann, Steven F. Roth, Morgan Hankins, and Maxine Eskenazi for helping evaluate the reading coach; Brian Milnes, Chetan Trikha, Stephen Reed, Scott Dworkis, Bryan Nagy, and David Sell for programming; *Weekly Reader*, Maxine Eskenazi, and Jennifer Gutwaks for text materials; and students, educators, and parents for Tutor tests in our lab and at Fort Pitt Elementary.

References

G. Aist. 1997. Challenges for a mixed initiative spoken dialog system for oral reading tutoring. AAAI 1997 Spring Symposium on Computational Models for Mixed Initiative Interaction.

J. R. Anderson, A. T. Corbett, K. Koedinger, and R. Pelletier. 1995. Cognitive tutors: Lessons learned. The Journal of Learning Sciences, 4,167-207.

J. Bernstein and D. Rtischev. 1991. A voice interactive language instruction system. Proceedings of the Second European Conference on Speech Communication and Technology (EUROSPEECH91). Genova, Italy.

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

M. E. Curtis. 1980. Development of components of reading skill. *Journal of Educational Psychology*, 72(5), 656-669.

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

I. Kantrov. 1991. Talking to the Computer: A Prototype Speech Recognition System for Early Reading Instruction, Technical Report, 91-3, Center for Learning, Teaching, and Technology, Education Development Center, 55 Chapel Street, Newton, MA 02160.

J. J. L'Allier. 1980. An evaluation study of a computer-based lesson that adjusts reading level by monitoring on task reader characteristics. Doctoral dissertation, University of Minnesota. Unpublished doctoral dissertation.

Lepper, M. R., Woolverton, M., Mumme, D. L., and Gurtner, J. 1993. Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors. In S. P. Lajoie and S. J. Derry (Ed.), *Computers as Cognitive Tools*. Hillsdale, NJ: Erlbaum.

Levy, B.A., Nicholls, A. and Kohen, D. 1993. Repeated readings: Process benefits for good and poor readers. *Journal of Experimental Child Psychology*, 56, 303-327.

I. Lundberg and A. Olofsson. Summer-Fall 1993. Can computer speech support reading comprehension**?** *Computers in Human Behaviour*, 9(2-3), 283-93.

J. Martin & K. VanLehn. 1995. A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman & R. L. Brennan (Eds.) Cognitively Diagnostic Assessment. Hillsdale, NJ: Erlbaum. pp. 141-165.

G. W. McConkie and D. Zola. 1987. Two examples of computer-based research on reading: Eye movement tracking and computer aided reading. In D. Reinking (Eds.), *Computers and Reading: Issues for Theory and Practice*. New York: Teachers College Press.

K. McCoy and D. Pany. 1986. Summary and analysis of oral reading corrective feedback research. *The Reading Teacher*, 39, 548-555.

J. Mostow, S. Roth, A. G. Hauptmann, and M. Kane. August 1994. A Prototype Reading Coach that Listens. Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94). Seattle, WA, American Association for Artificial Intelligence.

J. Mostow, A. Hauptmann, and S. Roth. 1995. Demonstration of a Reading Coach that Listens. Proceedings of the Eighth Annual Symposium on User Interface Software and Technology. Pittsburgh, PA: Sponsored by ACM SIGGRAPH and SIGCHI in cooperation with SIGSOFT.

M. Phillips, M. McCandless, and V. Zue. September 1992. Literacy Tutor: An Interactive Reading Aid Technical Report, Spoken Language Systems Group, MIT Laboratory for Computer Science, MIT.

R. K. Olson and B. Wise. 1987. Computer speech in reading instruction. In D. Reinking (Eds.), *Computers and Reading: Issues for Theory and Practice*. New York: Teachers College Press.

Olson, R.K. and Wise, B.W. 1992. Reading on the computer with orthographic and speech feedback: An overview of the Colorado remedial reading project. *Reading and Writing: An Interdisciplinary Journal*, 4, 107-144.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. 1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56-115.

D. Reinking and L. Bridwell-Bowles. 1991. Computers in reading and writing. In R. Barr, M. L. Kamil, P. B. Mosenthal, and P. D. Pearson (Eds.), *Handbook of reading research*. White Plains, NY: Longman.

P. Reitsma. 1988. Reading practice for beginners: Effects of guided reading, reading-while-listening, and independent reading with computer-based speech feedback. *Reading Research Quarterly*, 23(2), 219-235.

M. Russell, C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker. 1996. Applications of automatic speech recognition to speech and language development in young children. Proceedings of the Fourth International Conference on Spoken Language Processing. Philadelphia PA.

G. D. Spache. 1981. *Diagnostic Reading Scales*. Del Monte Research Park, Monterey, CA 93940: CTB Macmillan/McGraw-Hill.