Evaluating and Improving Real-Time Tracking of Children's Oral Reading

Yuanpeng Li and Jack Mostow
School of Computer Science, Carnegie Mellon University
mostow@cs.cmu.edu, yuanpeng.li@cs.cmu.edu

Abstract

The accuracy of an automated reading tutor in tracking the reader's position is affected by phenomena at the frontier of the speech recognizer's output as it evolves in real time. We define metrics of real time tracking accuracy computed from the recognizer's successive partial hypotheses, in con trast to previous metrics computed from the final hypothesis. We analyze the resulting considerable loss in real time ac curacy, and propose and evaluate a method to address it. Our method raises real time accuracy from 58% to 70%, which should improve the quality of the tutor's feedback.

Introduction

Automated reading tutors display text and listen to children read it aloud (Adams, 2006; Hagen, Pellom, and Cole, 2007; Mostow and Aist, 1999). They use automatic speech recognition (ASR) for at least two purposes. By aligning the ASR output with the text, a tutor tracks the reader's position in the text. By comparing each text word with the hypothesized word aligned against it, a tutor detects oral reading miscues.

Conventional metrics of ASR accuracy, such as word error rate (Hagen, et al., 2007), do not fit this task (Mostow, 2006), because it does not require *transcribing* miscues, just *detecting* them (Mostow et al., 1994).

Previous evaluations of ASR accuracy in reading tutors (other than of word error rate) have focused on miscue detection (Adams, 2006; Banerjee, Beck, and Mostow, 2003; Mostow et al., 1993; Mostow, et al., 1994; Tam et al., 2003) or accuracy in measuring oral reading rate (Balogh et al., 2007; Jian and Jianqiang, 2010).

Tracking accuracy is also important, but we have found almost no published evaluations of it, other than by Rasmussen *et al.* (2011). A near-exception occurs in tracking a vocal performance of a known score (Grubb and Dannenberg, 1997), the most closely related task in the area of automated score-following (Orio, Lemouton, and Schwarz, 2003) because it involves real-time speech recognition of a known text.

One reason for the importance of accuracy in tracking oral reading is that tracking accuracy is a limiting factor on the accuracy of miscue detection. If a tutor is wrong about which word the child is on, deciding whether it was read correctly is irrelevant, and liable to cause false rejection of the word the child was actually reading. Another reason involves graphical feedback provided by the tutor (Sitaram et al., 2011). Displaying the wrong position could be confusing or distracting, and even disrupt reading by causing children to lose their place. If children rely on the tutor to acknowledge each read word by updating the display, then failure to advance to the next word will make the tutor appear to have rejected the current word.

Rasmussen *et al.* (2011) defined *perceived tracking ac curacy* over the course of an utterance by how often the displayed position matched the true position, but measured it based on the final ASR hypothesis at the end of utterance. However, real-time tracking is based on the partial hypotheses emitted by the ASR during the utterance, which are not always initial subsequences of the final hypothesis. Partial hypotheses are especially unstable at the frontier – precisely the region most important in computing position. Two phenomena at this frontier increase tracking error.

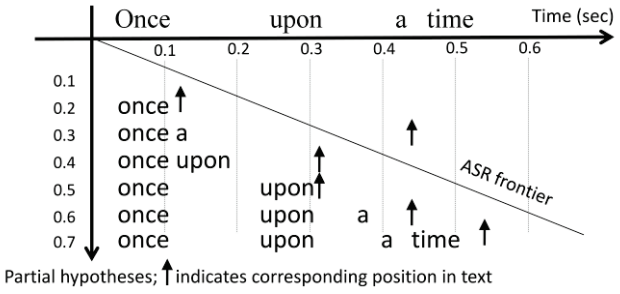

Figure 1. Example utterance

One source of error is ASR delay in hypothesizing a word in order to postpone the penalty for adding a word. In Figure 1, "Once" starts before t = 0.1, but the ASR does not hypothesize it until t = 0.2, when it generates the partial hypothesis "Once." At t = 0.3, it generates "Once a," which if correct would mean the reader had skipped a

word. Subsequent partial hypotheses can correct such tracking errors based on additional evidence that becomes available only as the frontier moves further on in the speech signal – but these corrections come only after the fact. Thus there is a tradeoff between accuracy and timeliness: waiting for more input can fix errors in tracking where the reader was at time t – but not until time t + lag.

Another source of errors is hypothesizing a word prematurely, for example hallucinating a word because it is strongly predicted by the language model or resembles background noise at that point in the input signal. For instance, the partial hypothesis at $t = 0.4$ under-estimates the start time of "upon," which actually starts after $t = 0.2$. In this example, hallucination merely causes the ASR to prematurely hypothesize the correct word and advance the estimated position. However, the ASR can also prematurely hypothesize words the reader did not say, e.g. "a" at $t = 0.3$.

Consequently, as the ASR processes oral reading, it tends to repeat the following cycle:

1. **Detect:** As the recognizer processes a spoken word, it searches for words that might match it. Once it accumulates enough acoustic evidence, the best-matching word starts appearing in the partial hypotheses it outputs.
2. **Hallucinate:** When this word first appears, it may be based on fragmentary evidence. In some cases it's incorrect, because the correct word is not the best match based on the evidence so far.
3. **Recognize prematurely:** Even when correct, it is often premature, that is, the reader has not yet finished uttering the word, but enough to make it the best match. Premature recognition of a word can cause the tutor to accept it even if the student omits the rest of the word – "a common short-cut among weaker readers that interferes with sight word and vocabulary growth (Adams, 1990) [and] a common speech habit within African American Vernacular English that interferes with reading acquisition" (Adams, 2011).
4. **Recognize correctly:** As the ASR continues to process the input speech, it eventually reaches the end of the spoken word. If this word appears at the end of the partial hypothesis, its end time now matches the speech.
5. **Procrastinate:** However, that is not the end of the story. As the ASR processes the ensuing speech, it keeps extending the word's recognized end time past its true end time. The higher the word insertion penalty imposed for hypothesizing an additional word, the longer the ASR puts off doing so, and the longer the recognized end time stretches past the actual end of the word.
6. **Recover:** This procrastination continues until the word's acoustic mismatch with the extended interval of speech becomes poor enough to rate some other hypothesis higher, such as appending an additional word or fragment to match the beginning of the next spoken word. At this point, the recognized word's aligned end time typically snaps back to (or near) the spoken word's actual end time. However, at this point the cycle is already repeating, starting at phase 1 for the next word.

Metrics of real-time accuracy

To measure tracking accuracy, we use the reader's actual position in the text to evaluate the position computed by the reading tutor after each partial ASR hypothesis. However, this computation may lag behind real-time.

We compare the estimated and true positions at time t of the word the reader is reading or just read. $PHypPos_u(t, lag)$ is the estimated position at time t based on the partial hypothesis output by the ASR at time $t + lag$ for the utterance u. $RefPos_u(t)$ is the true position at time t in utterance u, based on the time-aligned reference transcript. We define the distance between them as:

$$Dist_u(t, lag) = PHypPos_u(t, lag) - RefPos_u(t)$$

We measure tracking error by how often this distance exceeds some threshold:

$$Error_u(lag, \delta) = \#\{t \in T_u : |Dist_u(t, lag)| > \delta\}$$

Here T_u is the time interval spanned by the utterance u, and # *measures* the size of a set or time interval. The number and duration of utterances vary by student, and we normalize tracking error for student s by the total duration of that student's reading:

$$Error_s(lag, \delta) = \frac{\sum_{u \in U(s)} Error_u(lag, \delta)}{\sum_{u \in U(s)} \#T_u}$$

Here $U(s)$ is the set of utterances by student s. Next we compute the average error over the set S of students:

$$Error(lag, \delta) = \frac{\sum_{s \in S} Error_s(lag, \delta)}{\#S}$$

We define two kinds of accuracies. Exact Accuracy measures the percentage of time during which the estimated position matches the actual position:

$$ExactAccuracy(lag) = 1 - Error(lag, 0)$$

Near Accuracy measures the percentage of time where they differ by at most one word:

$$NearAccuracy(lag) = 1 - Error(lag, 1)$$

Results

The data come from 40 girls and 40 boys selected randomly from children for whom we have human transcripts of computer-assisted oral reading. The data set contains 2,033 transcribed utterances totaling 2.95 hours.

Table 1 and Figure 2 show how accuracy improves as lag increases. Figure 2 plots Exact Accuracy (the solid line) and Near Accuracy (the dotted line) as functions of lag. The thin dotted lines show 95% confidence intervals, computed as ± 1.96 times standard error of per-student mean accuracy. Treating the student as the unit of analysis controls for within-student statistical dependencies and avoids skewing results toward students who read more.

Figure 3 shows the distribution of distances from correct to estimated positions: 0, -1 (one word behind), +1 (one word ahead), and off by more than one word. Exact Accuracy is the proportion of 0's, and rises with lag.

Table 1. Tracking accuracy

	Overall		On task speech	
Lag	**Exact**	**Near**	**Exact**	**Near**
0 (baseline)	58.1%	81.6%	63.4%	86.8%
0.1 sec	66.1%	82.6%	72.6%	88.0%
0.2 sec	70.2%	83.6%	77.3%	89.1%
0.5 sec	72.9%	85.2%	80.4%	91.0%
1.0 sec	73.9%	86.3%	81.4%	92.2%
1.5 sec	74.4%	86.7%	81.9%	92.5%
2.0 sec	74.6%	86.9%	82.1%	92.6%
Final Hyp	75.6%	87.3%	83.3%	92.8%

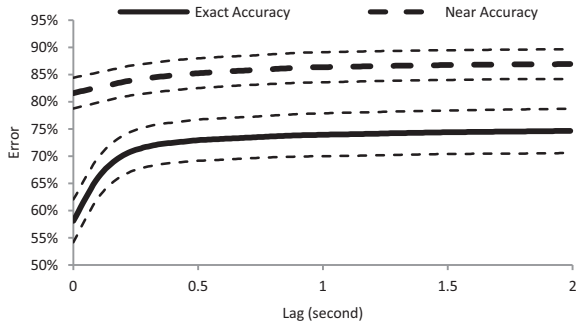

Figure 2. Accuracy vs. lag

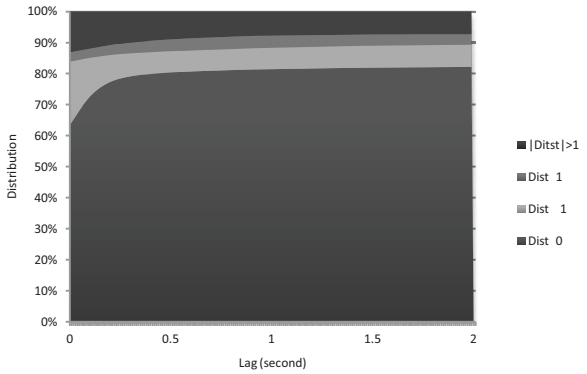

Figure 3. Distances from correct to estimated position

As lag increases from zero to 0.2 seconds, Exact Accuracy rises from 58.1% to 70.2%, and Near Accuracy from 81.6% to 83.6%. As lag continues to increase, so does accuracy, but more slowly. One explanation is that different types of errors are reduced during different phases.

The sharp increase in Exact Accuracy as lag increases from zero to about 0.2 seconds is due primarily to a reduction in the proportion of -1, corresponding to ASR delay in hypothesizing the next word until it recognizes the end of the word. Near Accuracy allows distance of -1, so this reduction does not affect it, which is why it improves more slowly than Exact Accuracy during the first phase.

In the second phase, words are hypothesized but the evidence is still sometimes insufficient for correct recognition. The resulting incorrect hypotheses can take the tracker anywhere in the sentence, which is often more than one word away from the reader's true position. As lag increases, the accumulating evidence enables the recognizer to resolve more such errors. Reduction of their proportion drives most of the accuracy improvement during this phase.

Lag increases until it culminates in the final hypothesis for the utterance, which takes advantage of maximum evidence, and hence tends to be the most accurate. Offline tracking as in Rasmussen et al.'s (2011) analysis is based on this final hypothesis, leading it to over-estimate real-time tracking accuracy.

Children's oral reading contains off-task speech (chatting, humming, etc.), during which the reader's position in the text does not advance (Chen, 2012). 17.6% of the speech in our data set is off-task. Figure 4 and Figure 5 show Exact Accuracy and Near Accuracy, respectively, disaggregated by whether the speech is on-task or off-task.

Both Exact Accuracy and Near Accuracy are much lower for off-task speech, because the ASR expects to hear the words in the current sentence, in the same order – not the words in off-task speech. Sentence words, especially high-frequency words like "I" and "the," do occur in off-task speech, but seldom two or more in sequence, let alone in the same order. Interpreting off-task speech as oral reading causes the ASR to jump around in the sentence to string together words it is listening for into whatever sequence most resembles the off-task speech. The ASR output causes tracking to jump around also, rather than stay in place as it should, and its accuracy suffers greatly as a consequence.

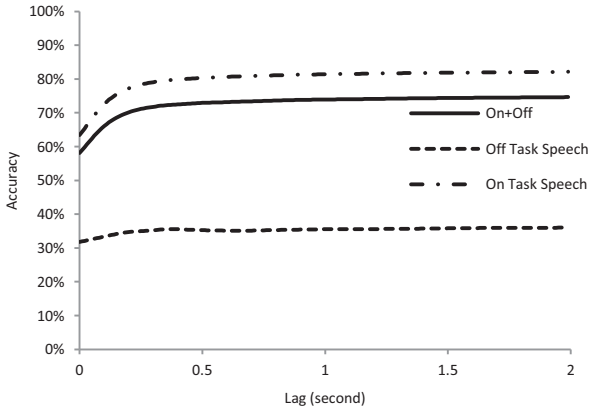

Figure 4. Exact Accuracy for on/off task speech

Conclusion

In this paper, we proposed a metric to evaluate real-time tracking accuracy of oral reading based on partial hypotheses output by ASR. We found that a lag as brief as 0.2 seconds can improve tracking accuracy dramatically. The

HCI implications of this finding are significant for designing the interface of a reading tutor, specifically the visual display of estimated position, because a 0.2 second lag in updating such a display would be fast enough for the update to seem prompt. Even a normal speaking rate of 180 words per minute takes a third of a second to speak a word. Thus even for a fluent oral reader, the tutor would update the display to credit a word before the student finished speaking the next word. In contrast, a longer lag might lead the student to reread a word to repair a perceived error, whether in the student's reading or the tutor's hearing.

We also found that tracking accuracy is much lower during off task speech. This result implies that filtering out off-task speech (Chen and Mostow, 2011) or suspending tracking when off-task speech is detected could further improve accuracy in real-time tracking of oral reading.

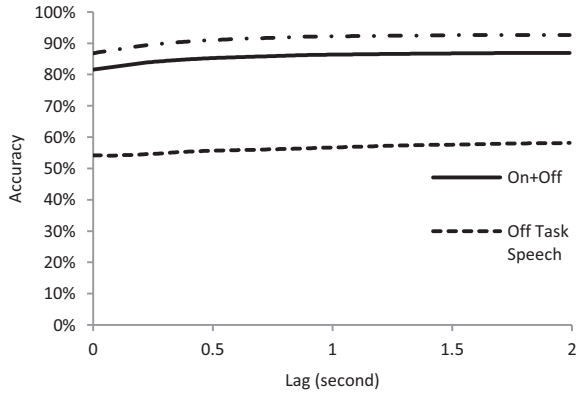

Figure 5. Near Accuracy for on/off task speech

References

Adams, M. J. 1990. *Beginning to Read: Thinking and Learning about Print*. Cambridge, MA: MIT Press.

Adams, M. J. (2006). The promise of automatic speech recognition for fostering literacy growth in children and adults. In M. McKenna, L. Labbo, R. Kieffer & D. Reinking (Eds.), *International Handbook of Literacy and Technology* (Vol. 2, pp. 109 128). Hillsdale, NJ: Lawrence Erlbaum Associates.

Adams, M. J. 2011. Technology for Developing Children's Language and Literacy: Bringing Speech Recognition to the Classroom. New York, NY: The Joan Ganz Cooney Center at Sesame Workshop.

Balogh, J., Bernstein, J., Cheng, J., and Townshend, B. 2007. Automatic evaluation of reading accuracy: assessing machine scores. In *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE)*, 112 115. Farmington, PA.

Banerjee, S., Beck, J. E., and Mostow, J. 2003. Evaluating the effect of predicting oral reading miscues. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 3165 3168. Geneva, Switzerland.

Chen, W. (2012). *Detecting Off task Speech.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

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.

Grubb, L. and Dannenberg, R. B. 1997. A Stochastic Method of Tracking a Vocal Performer. In *Proceedings of the International Computer Music Conference*, 301 308. Thessaloniki, Greece.

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

Jian, C. and Jianqiang, S. 2010. Towards accurate recognition for children's oral reading fluency. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, 103 108.

Mostow, J. 2006. Is ASR accurate enough for automated reading tutors, and how can we tell? In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 ICSLP), Special Session on Speech and Language in Education*, 837 840. Pittsburgh, PA.

Mostow, J. and Aist, G. S. 1999. Giving help and praise in a reading tutor with imperfect listening because automated speech recognition means never being able to say you're certain. *CALICO Journal, 16*(3), 407 424.

Mostow, J., Hauptmann, A. G., Chase, L. L., and Roth, S. 1993. Towards a reading coach that listens: automated detection of oral reading errors. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93)*, 392 397. Washington, DC.

Mostow, J., Roth, S. F., Hauptmann, A. G., and Kane, M. 1994. A prototype reading coach that listens [AAAI 94 Outstanding Paper]. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 785 792. Seattle, WA.

Orio, N., Lemouton, S., and Schwarz, D. 2003. Score following: state of the art and new developments. In *Proceedings of the 2003 conference on New interfaces for musical expression*, 36 41. Montreal, Quebec, Canada.

Rasmussen, M. H., Mostow, J., Tan, Z. H., Lindberg, B., and Li, Y. 2011. Evaluating Tracking Accuracy of an Automatic Reading Tutor. In *SLaTE: ISCA (International Speech Communication Association) Special Interest Group (SIG) Workshop on Speech and Language Technology in Education*, Venice, Italy.

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.

Tam, Y. C., Mostow, J., Beck, J., and Banerjee, S. 2003. Training a Confidence Measure for a Reading Tutor that Listens. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 3161 3164. Geneva, Switzerland.