Is ASR accurate enough for automated reading tutors, and how can we tell?

Jack Mostow

Project LISTEN, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
Soliloquy Learning, Waltham, MA, USA
mostow@cs.cmu.edu

ABSTRACT

We discuss pros and cons of several ways to evaluate ASR accuracy in automated tutors that listen to students read aloud. Whether ASR is accurate enough for a particular reading tutor function depends on what ASR-based judgment it requires, the visibility of that judgment to students and teachers, and the amount of input speech on which it is based. How to tell depends on the purpose, criterion, and space of the evaluation.
Index Terms: speech recognition, evaluation, reading tutors

1. INTRODUCTION

Automated continuous speech recognition (ASR) is used in a growing number of systems to help students learn to read: from past experiments [1, 2] to prototypes being developed [3] to tutors used daily in schools [4] to commercial products [5]. ASR is also used to evaluate pronunciation [6].

The ultimate criterion for evaluating tutors is their impact on student learning, measured by controlled studies of students' gains from pre- to post-test, compared to alternative treatments [4, 5, 7-11]. Such studies require months of data and cannot be rerun off-line. Their results depend on the population sample, the entire design of the tutor, how much it is used, and the control(s) against which it is compared. They evaluate the tutor's overall impact, not its ASR accuracy.

Is ASR accurate enough to use in an automated reading tutor? This question is too broad as phrased, because the answer depends on how the tutor uses ASR. Unlike conventional ASR used to transcribe unknown speech, reading tutors know the text the student is supposed to read. They use ASR to track the reader's position in the text, detect miscues, and measure word reading times. This paper discusses the evaluation of ASR accuracy for those functions.

Future tutors may also use ASR to engage in spoken dialogue about the text [12]. The less constrained the input, the more the task resembles spontaneous speech recognition from the ASR's point of view, in which case conventional word error rate (WER) or semantic error rate may be a useful metric to evaluate its accuracy. The more constrained the correct student responses in such dialogue, the more closely it resembles oral reading from the ASR's point of view.

A useful metric of ASR accuracy should match the function for which ASR is used. It should apply economically not only to the original tutor sessions but also to re-recognizing the recorded speech with modified ASR. It should cope gracefully with vocabulary mismatch between how ASR and human transcribers represent oral reading. A general metric should be able to compare ASR accuracy across different tutors, rather than be application-specific.

How to evaluate depends on the purpose of evaluation. One purpose is to decide whether ASR is accurate enough to support a given use: is Accuracy(ASR, data) > "OK"? Another purpose is to compare alternative ASR methods: is Accuracy(ASR1, data) > Accuracy(ASR2, data)? For example, is one recognizer significantly better than another? Does a proposed change actually improve ASR accuracy? Which parameter values are optimal? A third purpose is to compare data sets in order to quantify how they differ in difficulty, and to help understand why: when is Accuracy(ASR, data1) > Accuracy(ASR, data2)? A closely related purpose is to understand sources of ASR error by disaggregating speech within the same data set.

How good is good enough? The criterion depends on the tutorial judgments to be based on ASR. Judgments need to be more accurate if they are visible to students and teachers than if used just to guide covert tutorial decisions. Tutorial judgments are more robust to random ASR errors if aggregated over more than one spoken word. Averaging N independent estimates reduces error by a factor of $\sqrt{N}$.

In what space should evaluation be performed? *Text space* consists of the text words to be read. *Speech space* is the spoken sequence of words heard by ASR or a human transcriber. *Time domain* aligns spoken words to times.

This paper discusses ASR accuracy for three functions: tracking the reader's position in a text (Section 2), detecting reading mistakes (Section 3), and measuring word reading times (Section 4). We discuss metrics for various purposes, criteria, and spaces. Section 5 concludes.

2. TRACKING ACCURACY

Accuracy improvement is our main purpose in measuring how well Project LISTEN's Reading Tutor tracks a reader through a sentence. As [2] details, the tutor aligns real-time ASR output to the text to locate the reader's current position, identify which text words the reader tries to read, and detect when the reader skips a word or reaches the end of the sentence. Thus the accuracy criterion for tracking is how well ASR guides these individual tutoring decisions.

We evaluate tracking accuracy in speech space on a transcribed corpus of oral reading recorded by the Reading Tutor. First we align each transcript to the text to compute the reader's true path through the text. We align the ASR output to the text to find the path it "heard." We represent each path as a sequence of word positions in the text, marked by whether that word was read correctly. We then

compute the rate of the tracking errors where the ASR-based path does not follow the transcript-based path.

In regions where off-task speech interrupts reading, the aligned word position is not very meaningful. An off-task region manifests as a sequence of word positions marked as not read correctly, except for occasional words like *I* or *the* that happen to match the text. We define *deviation length* as the number of words in such a sequence. Deviations longer than 5 words generally consist of off-task speech rather than a series of attempts to read words.

To measure tracking error separately for correct reading, off-task speech, and misreading, we disaggregate tracking error by deviation length. We find that tracking accuracy is highest for correct reading. Tracking is poor for off-task reading, but doesn't hamper tutoring. But a tutor must track misreading very accurately in order to give immediate corrective feedback on the right word.

3. MISCUE DETECTION

One purpose of evaluating the accuracy of ASR in detecting reading mistakes is to improve it. Another is to characterize what reading tutors should or should not count on ASR to do.

What criterion should define what to count as a mistake? One criterion is any deviation from perfect reading, according to an orthographic and/or phonetic transcript. Transcription standards must specify how to classify a word as correct, an acceptable dialectal variant, or a mistake. However, given well-defined standards and adequate training, transcribers can achieve high inter-rater reliability. The resulting transcripts are amenable to useful analyses, such as training and evaluating predictive models of phoneme-level errors [13].

However, not all deviations from perfect reading matter. A more application-oriented criterion is whether they do. One such criterion is whether a miscue is serious enough to threaten comprehension [2]. A similar criterion is if a tutor should intervene [5]. Such criteria fit the tutorial decisions that ASR supports. They distinguish reading mistakes from dialect phenomena, and ignore mistakes too minor to matter. However, these criteria are subjective in nature, relying as they do on individual judgments of which mistakes matter. Teachers disagree, so their inter-rater reliability is limited [5].

The costs of errors in detecting reading mistakes have an interesting asymmetry. There is a *motivational* cost for false alarms, that is, words read correctly by the student but rejected by the tutor. Experience with deployed reading tutors suggests that children tolerate a false alarm rate of a few percent well enough to use automated reading tutors for a whole school year, but get frustrated by repeated false alarms.

Conversely, there is a *cognitive* cost for undetected miscues, that is, words misread or omitted by the student but accepted by ASR. This type of ASR error can deprive the tutor of opportunities to remediate student mistakes. It is not clear how to quantify the cost of such errors, especially since other tutorial responses may provide relevant feedback. For instance, a reading mistake is often accompanied by halting, disfluent reading, to which Project LISTEN's Reading Tutor responds by rereading the sentence even if it did not detect the mistake [14]. This response may serve as corrective feedback, provided the student attends to the corrected word.

Judgments about individual words require the greatest accuracy. The need for accuracy is especially great for overt tutorial judgments visible to students and teachers. For example, if a tutor explicitly announces whether a student read a word correctly, or colors each word red that it thinks the student misread, then ASR errors place its credibility at risk. Even a single ASR error per 100 words amounts to multiple errors per session. It seems unreasonable to expect students or teachers to trust a tutor that is wrong so often.

The need for accuracy is somewhat lower for *implicit* judgments about individual words – judgments that are not explicitly announced or displayed, but that guide tutorial decisions, such as whether to give help on a word.

Aggregating over a student's successive attempts to read a word can improve the accuracy of such decisions. Beck *et al.* [15] used ASR of each such attempt to update the probability of the student knowing that particular word. This aggregated estimate scored the next attempt more accurately than ASR of the attempt itself.

Likewise, aggregating over different students and words enables statistically reliable comparisons of the efficacy of different types of tutorial assistance on words, based on ASR judgments of students' performance when they next encounter those words [16].

There is more than one space in which to evaluate ASR accuracy, each with advantages and disadvantages.

"Text space" evaluation of ASR measures how accurately it classifies each word of text as correct, misread, or omitted. Mostow *et al.* [2] simply classified text words as (ultimately) read correctly or not.

One advantage of text space is that the set of tokens in a given text is well-defined, and invariant across ASR runs. Text space evaluation is also application-appropriate. It measures how well the tutor detects the mistakes that matter – namely, the mistakes that the student does not self-correct, and which it may therefore be appropriate for the tutor to remediate.

A disadvantage of text space evaluation is sparse data on missed words, especially in a tutor that gives help on demand. A standard authority [17] considers text to exceed a reader's "frustration level" if the reader makes more than one miscue per 10 text words – including not just missed words but hesitations, repetitions, and self-corrections. Consequently missed words are much rarer than correct words in text space, leading to poor estimates of how accurately they are detected.

"Speech space" evaluation classifies transcribed or recognized words instead of text words. Speech space has more examples of misread or omitted words, because it includes reading mistakes self-corrected by the reader.

A disadvantage of speech space is that its unit of analysis may be ill-defined. For example, if the reader haltingly and errorfully sounds out a word, it may be hard to decide which sounds or sequences of sounds to count as spoken words, let alone mark them automatically in human transcripts or ASR output.

"Time domain" evaluation compares time-aligned ASR output to a time-aligned transcript of what the reader said. It credits the ASR for accepting a correctly read word only if the ASR heard the word at the interval in the speech signal where the reader actually spoke it. This more stringent criterion avoids crediting the ASR for hallucinating correct reading due to its strong language

model of the text [2]. The resulting more realistic evaluation can give a clearer picture of how accurately the ASR is really behaving. However, time domain is vulnerable to transcript errors and misalignment of transcript to text. Forced alignment of the transcript to the student's recorded oral reading is imperfect, but manual time alignment is expensive.

Given a criterion for what to detect, various metrics quantify miscue detection accuracy – but some are flawed.

Word error rate (WER) measures how well ASR recognizes what the reader said, independent of the text. Hagen *et al.* [18] claim "word error rate calculations using the widely accepted NIST scoring software provides the most widely accepted, easy to use and highly valid metric."

However, WER does not directly measure how accurately ASR performs the tracking, detection, or timing functions it serves in a reading tutor. In particular, WER gives ASR zero credit for *detecting* reading mistakes unless it correctly *recognizes* the exact miscue the reader uttered – which is both rare and unnecessary for tutorial intervention. For example, suppose a reader misreads *elegant* as *elephant*, which is not in the text. If ASR outputs *and of that* instead, it will detect the miscue – yet incur at least as high a WER penalty as if it accepts *elegant* as read correctly. Hence using WER to optimize ASR parameters penalizes detection. The resulting ASR configuration tends to classify miscues as read correctly because WER provides no incentive to reject them.

Finally, WER is vulnerable to vocabulary mismatch between ASR and human transcripts. For instance, a transcript may use words to represent oral reading miscues that ASR uses other symbols to represent. Phoneme error rate (PER) avoids this vocabulary mismatch problem by using phonemes as a common representation. However, PER still measures accuracy of recognition rather than detection.

The overall percentage of words correctly classified tells little, as it typically rises with the percentage correctly read. Some researchers report rates of false positives (words falsely accepted) and false negatives (words falsely rejected). These rates measure how often ASR is wrong in each way, but they still vary with the percentage read correctly – and they conceal whether ASR is any better than random. For instance, say the student misreads 5 of 100 words, and ASR rejects 5 *other* words but accepts the rest as correct. The FP and FN rates are each only 5%, which sounds good. Yet detection of misread words is *worse* than random here. Better just to accept all words, achieving 5% FP and 0% FN!

We find it clearer to compute separate error rates for different categories, such as correct, omitted, or misread. We define such rates in both text and speech space. The denominator is the number of words in that category. The numerator is how many of them are misclassified by ASR.

In particular, the *false alarm* rate is the percentage of correctly read words rejected by ASR. The *miscue detection* rate is the percentage of misread words rejected. Together, the false alarm and miscue detection rates give a useful characterization of a tutor's ASR accuracy, not artificially skewed by the proportion of correct words. However, they specify only a single point on a tradeoff curve. To quantify the accuracy of a confidence score, we plot the ROC curve of miscue detection rate versus false alarm rate as the threshold score to accept a word varies. The area under the ROC curve summarizes the ASR confidence score's overall accuracy.

4. WORD READING TIME

The time to read a word of text is a fine-grained indicator of reading proficiency and growth [19]. The purpose of evaluating how well ASR measures reading times is to validate and refine its ability to assess reading proficiency.

A direct criterion is agreement of time-aligned ASR output with a time-aligned human transcript, but alignment is expensive if manual, and imperfect if automated.

An indirect criterion is the ability to predict students' test scores from their distributions of word reading times. This approach bypasses the need for human transcripts by aggregating over many words of ASR output, which is less accurate than human transcripts but much more plentiful. Beck *et al.* [20] estimated each student's oral reading fluency by aggregating over the inter-word latencies preceding all the words read by the student within a time window of a few weeks. The resulting fluency estimates correlated well with paper tests of oral reading fluency.

We used this indirect criterion to choose among alternative ways to operationalize inter-word latency – an issue that not even perfect alignment would have resolved, because it was definitional. We simply picked the version that best predicted test scores. This criterion was based on ASR output, enabling us to use a massive quantity of oral reading instead of the small percentage transcribed by hand. Moreover, it matched a tutor function we wanted to support – namely, assessment of proficiency.

Aggregating over multiple students can further increase robustness to ASR errors. Even well-validated paper tests with high statistical reliability are subject to measurement errors at the level of individual students. Such errors may occasionally cause individual students' scores to decline from pre- to posttest even though their knowledge actually grew. However, measures unreliable at the level of individual students can nevertheless provide statistically solid grounds for conclusions about sufficiently large groups of students, as illustrated by the National Assessments of Academic Progress [21]. Likewise, some measures of reading performance based on noisy ASR may spuriously indicate declines over time for some individual students, yet provide reliable results when applied to a larger sample of students, such as a class or a reading group. For example, such aggregation might be used to estimate how many students know some word, or to compute the average fluency growth of a reading group.

In general, the amount of data required for a "good enough" aggregate judgment depends not only on the error of the measure, but also on the decision based on the judgment, and the costs of error. For example, estimating student reading proficiency to the nearest grade level is probably good enough to decide what level of material to read, especially because the availability of tutorial assistance on demand reduces the cost of misestimation. However, such a rough estimate is not accurate enough to monitor a student's weekly progress.

5. CONCLUSIONS

How accurate is ASR, and how can we tell? The general answer is "It depends" – in particular, on the function for

which ASR is used, the purpose and criterion for evaluation, and the space in which evaluation is done. Quantitative results vary among data sets, but qualitatively, ASR can:

- Track the reader's position in a sentence well enough to tell when the reader skips a word or finishes the sentence, but not always which word to correct when the reader misreads.
- Score reading well enough to avoid frustration and detect a portion of the miscues a human would correct, but not to tell students or teachers reliably which words are right or wrong.
- Measure aggregated word reading times well enough to estimate student reading level and report fluency growth.

Acknowledgments

This work was supported by the National Science Foundation, ITR/IERI Grant No. REC-0326153. Any opinions, findings, and 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 sponsor(s) or of the United States Government. Thanks to Joe Beck, Ravi Mosur, and Evandro Gouvea for their work and comments, and to schools for data.

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

1. Russell, M., C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker. Applications of automatic speech recognition to speech and language development in young children. *Proceedings of the Fourth International Conference on Spoken Language Processing* 1996. Philadelphia PA.
2. Mostow, J., S.F. Roth, A.G. Hauptmann, and M. Kane. A prototype reading coach that listens [AAAI-94 Outstanding Paper]. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 785-792. 1994. Seattle, WA: American Association for Artificial Intelligence.
3. Hagen, A., B. Pellom, and R. Cole. Children's Speech Recognition with Application to Interactive Books and Tutors. *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop* 2003. St. Thomas, USA.
4. Mostow, J., G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, M.B. Sklar, and B. Tobin. Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 2003. *29*(1): p. 61-117.
5. Adams, M.J. The promise of automatic speech recognition for fostering literacy growth in children and adults. In M. McKenna, et al., Editors, *International Handbook of Literacy and Technology*, 109-128. Lawrence Erlbaum Associates: Hillsdale, NJ, 2006.
6. Neumeyer, L., H. Franco, V. Digalakis, and M. Weintraub. Automatic scoring of pronunciation quality. *Speech Communication*, 2000. *30*(2-3): p. 83-93.
7. Cunningham, T. *The Effect of Reading Remediation Software on the Language and Literacy Skill Development of ESL Students.* Unpublished Master's thesis, University of Toronto, Toronto, Canada, 2006.
8. Mostow, J., G. Aist, C. Huang, B. Junker, R. Kennedy, H. Lan, D. Latimer, R. O'Connor, R. Tassone, B. Tobin, and A. Wierman. 4-Month evaluation of a learner-controlled Reading Tutor that listens. In V.M. Holland and F.N. Fisher, Editors, *Speech Technology for Language Learning*. Taylor & Francis in press.
9. Poulsen, R. *Tutoring Bilingual Students With an Automated Reading Tutor That Listens: Results of a Two-Month Pilot Study*. Unpublished Master's thesis, DePaul University, Chicago, IL, 2004.
10. Reeder, K., M. Early, M. Kendrick, J. Shapiro, and J. Wakefield. The Role of L1 in Young Multilingual Readers' Success With a Computer-Based Reading Tutor. *Fifth International Symposium on Bilingualism* 2005. Barcelona, Spain.
11. Mostow, J., G. Aist, J. Bey, P. Burkhead, A. Cuneo, B. Junker, S. Rossbach, B. Tobin, J. Valeri, and S. Wilson. Independent practice versus computer-guided oral reading: Equal-time comparison of sustained silent reading to an automated reading tutor that listens. *Ninth Annual Meeting of the Society for the Scientific Study of Reading* 2002. Chicago, Illinois.
12. Lee, K., A. Hagen, N. Romanyshyn, S. Martin, and B. Pellom. Analysis and Detection of Reading Miscues for Interactive Literacy Tutors. *20th International Conference on Computational Linguistics (Coling)* 2004. Geneva, Switzerland.
13. Banerjee, S., J. Beck, and J. Mostow. Evaluating the Effect of Predicting Oral Reading Miscues. *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 3165-3168. 2003. Geneva, Switzerland.
14. Mostow, J. and G. Aist. 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*, 1999. *16*(3): p. 407-424.
15. Beck, J.E., K.-m. Chang, J. Mostow, and A. Corbett. Using a student model to improve a computer tutor's speech recognition. *Proceedings of the AIED 05 Workshop on Student Modeling for Language Tutors, 12th International Conference on Artificial Intelligence in Education*, 2-11. 2005. Amsterdam.
16. Heiner, C., J.E. Beck, and J. Mostow. Improving the help selection policy in a Reading Tutor that listens. *Proceedings of the InSTIL/ICALL Symposium on NLP and Speech Technologies in Advanced Language Learning Systems*, 195-198. 2004. Venice, Italy.
17. Betts, E.A. *Foundations of Reading Instruction*. 1946, New York: American Book Company.
18. Hagen, A., B. Pellom, S.v. Vuuren, and R. Cole. Advances in Children's Speech Recognition within an Interactive Literacy Tutor. *HLT NAACL* 2004. Boston.
19. Mostow, J. and G. Aist. The sounds of silence: Towards automated evaluation of student learning in a Reading Tutor that listens. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, 355-361. 1997. Providence, RI: American Association for Artificial Intelligence.
20. Beck, J.E., P. Jia, and J. Mostow. Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning*, 2004. *2*(1-2): p. 61-81.
21. NCES. NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance. 2001, National Center for Education Statistics, United States Department of Education.