Analysis and Detection of Reading Miscues for Interactive Literacy Tutors

Katherine Lee, Andreas Hagen, Nicholas Romanyshyn, Sean Martin, Bryan Pellom
Center for Spoken Language Research
University of Colorado at Boulder
pellom@cslr.colorado.edu

Abstract

The Colorado Literacy Tutor (CLT) is a technology-based literacy program, designed on the basis of cognitive theory and scientifically motivated reading research, which aims to improve literacy and student achievement in public schools. One of the critical components of the CLT is a speech recognition system which is used to track the child's progress during oral reading and to provide sufficient information to detect reading miscues. In this paper, we extend on prior work by examining a novel labeling of children's oral reading audio data in order to better understand the factors that contribute most significantly to speech recognition errors. While these events make up nearly 8% of the data, they are shown to account for approximately 30% of the word errors in a state-of-the-art speech recognizer. Next, we consider the problem of detecting miscues during oral reading. Using features derived from the speech recognizer, we demonstrate that 67% of reading miscues can be detected at a false alarm rate of 3%.

1 Introduction

Pioneering research by MIT and CMU as well as more recent work by the IBM *Watch-me-Read* project has demonstrated human language technologies can play an effective role in systems designed to improve children's reading abilities (McCandless, 1992; Mostow et al., 1994; Zue et al., 1996). In CMU's *Project LISTEN*, for example, the tutor operates by prompting children to read individual sentences out loud. The tutor listens to the child using speech recognition and extracts features that can be used to detect oral reading miscues (Mostow et al., 2002; Tam et al. 2003). The most common miscues that children make while reading out loud are word substitutions, repetitions, and self-corrections with word omissions and insertions being less frequent (Fogarty et al. 2001).

Upon detecting such reading errors, the tutor must provide appropriate feedback to the child. While the type of feedback and level of feedback is the current subject of much debate within the research community, recent results have shown that automated reading tutors can improve student achievement (Mostow et al., 2003). In fact, providing real time feedback by simply highlighting words as they are read out loud is the basis of at least one commercial product today[1].

Cole et al. (2003) and Wise et al. (in press) describe a new scientifically-based literacy program, *Foundations to Fluency*, in which a virtual tutor—a lifelike 3D computer model—interacts with children in multimodal learning tasks to teach them to read. A key component of this program is the Interactive Book, which combines real-time multilingual speech recognition, facial animation, and natural language understanding capabilities to teach children to read and comprehend text. Within the context of this reading program, Hagen et al. (2003) demonstrated an initial speech recognition system that provides real-time reading tracking for children. This work was later extended by Hagen et al. (2004) to incorporate improved acoustic and language modeling strategies. When tested on 106 children (ages 9-11) who were asked to read one of a number of short age-appropriate stories, a final system word error rate of 8.0% was demonstrated.

While reporting raw word error rate is useful for comparison purposes to prior research, we point out that it does not provide any diagnostic information which can be used to understand factors that contribute to speech recognition error within such children's literacy tutor programs. Therefore, this paper extends our earlier work in two important ways. First, in order to understand *where* future improvements can be obtained, we provide a novel "event" labeling of our children's speech corpus and examine the performance of the current speech recognition system under each labeled event condition. Second, we describe the construction of an automated classifier which can detect reading miscues in children's speech.

This paper is organized as follows. First, Section 2 provides an introduction and overview of the Colorado Literacy Tutor project. Section 3 describes the audio corpus used in the experiments

[1] http://www.soliloquy.com

provided in this paper and Section 4 describes our baseline speech recognition system. Next, Section 5 describes the event labeling methodology and word error analysis under each labeled event condition. Finally Section 6 describes our initial work towards developing a system to detect reading miscues based on the output of our baseline speech recognition system. Conclusions and future work are outlined in Section 7.

2 The Colorado Literacy Tutor

The Colorado Literacy Tutor (CLT)[2] is a technology-based literacy program, designed on the basis of cognitive theory and scientifically motivated reading research, which aims to improve literacy and student achievement in public schools. The goal of the Colorado Literacy Tutor is to provide computer-based learning tools that will improve student achievement in any subject area by helping students learn to read fluently, to acquire new knowledge through deep understanding of what they read, to make connections to other knowledge and experiences, and to express their ideas concisely and creatively through writing. A second goal is to scale up the program to both state and national levels in the U.S. by providing accessible, inexpensive and effective computer-based learning tools.

The CLT project consists of four tightly integrated components: Managed Learning Environment, Foundational Reading Skills Tutors, Interactive Books, and Latent-Semantic Analysis (LSA)-based comprehension training (Steinhart 2001; Deerwester et al., 1990; Landauer and Dumais, 1997). A key feature of the project is the use of leading edge human communication technologies in learning tasks. The project has become a test bed for research and development of perceptive animated agents that integrate auditory and visual behaviors during face-to-face conversational interaction with human learners. The project enables us to evaluate component technologies with real users—students in classrooms—and to evaluate how the technology integration affects learning using standardized assessment tools.

Within the CLT, Interactive Books are the main platform for research and development of natural language technologies and perceptive animated agents. Figure 1 shows a page of an Interactive Book. Interactive Books incorporate speech recognition, spoken dialogue, natural language processing, and computer animation technologies to enable natural face-to-face conversational interaction with users. The integration of these technologies is performed using a client-server architecture that provides a platform-independent user interface for Web-based delivery of multimedia learning tools. Interactive Book authoring tools are designed for easy use by project staff, teachers and students to enable authors to design and format books by combining text, images, videos and animated characters. Once text and illustrations have been imported or input into the authoring environment, authors can orchestrate interactions between users, animated characters and media objects. Developers can populate illustrations (digital images) with animated characters, and cause them to converse with each other, with the user, or speak their parts in the stories using naturally recorded or synthetic speech. A mark up language enables authors to control characters' facial expressions and gestures while speaking. The authoring tools also enable authors to pre-record sentences and/or individual words in the text as well as utterances to be produced by animated characters during conversations.

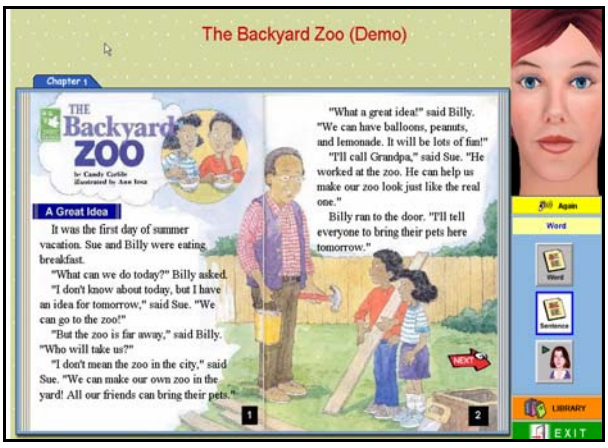


Figure 1: An example interactive book

Interactive Books enable a wide range of user and system behaviors. These include having the story narrated by animated characters, having conversations with animated characters in structured or mixed-initiative dialogues, having the student read out loud while words are highlighted, enabling the student to click on words to have them spoken by the agent or to have the agent interact with the student to sound out the word, having the student respond to questions posed by the agent either by clicking on objects in images or saying or typing responses, and having the student produce typed or spoken story summaries which can be analyzed for content using natural language processing techniques.

[2] http://www.colit.org

3 CU Children's Read Story Corpus

Within the context of the CLT project, we have collected a corpus of audio data consisting of read stories spoken by children. Known as the *CU Children's Read Story Corpus*[3], the data currently contains speech and associated word-level transcriptions from 106 children who were asked to read a short age-appropriate story and to provide a spontaneous spoken summary of the material. In addition, each child was prompted to read 25 phonetically balanced sentences for future use in exploring strategies for speaker adaptation.

The data were collected from native English speaking children in the Boulder Valley School District (Boulder, Colorado, USA). We have initially collected and transcribed stories from children in grades 3, 4, and 5 (grade 3: 17 speakers, grade 4: 28 speakers, grade 5: 61 speakers). The data were originally collected in a quiet room using a commonly available head-mounted microphone. The current 16 kHz sampled corpus consists of 10 different stories. Each story contains an average of 1054 words (min 532 words / max 1926 words) with an average of 413 unique words per story. Note that while each story is accompanied by a spontaneous summary produced by the child, we do not consider those data for this paper.

4 Baseline Speech Recognition System

The CLT uses the SONIC speech recognition system as a basis for providing real-time recognition of children's speech (Pellom, 2001; Pellom and Hacioglu, 2003; Hagen et al. 2004)[4]. The recognizer implements an efficient time-synchronous, beam-pruned Viterbi token-passing search through a static re-entrant lexical prefix tree while utilizing continuous density mixture Gaussian Hidden Markov Models (HMMs). The recognizer uses PMVDR cepstral coefficients (Yapanel and Hansen, 2003) as its feature representation. Children's acoustic models were estimated from 46 hours of audio from the *CU Read and Prompted Children's Speech Corpus* (Hagen et al., 2003)[5] and the OGI Kids' speech corpus (Shobaki et al., 2000).

During oral reading, the speech recognizer models the story text using statistical n-gram language models. This approach gives the recognizer flexibility to insert/delete/substitute words based on acoustics and to provide accurate confidence information from the word-lattice. The recognizer receives packets of audio and automatically detects voice activity. When the child speaks, the partial hypotheses are sent to a reading tracking module. The reading tracking module determines the current reading location by aligning each partial hypothesis with the story text using a Dynamic Programming search. In order to allow for skipping of words or even skipping to a different place within the text, the search finds words that when strung together minimize a weighted cost function of adjacent word-proximity and distance from the reader's last active reading location. The Dynamic Programming search additionally incorporates constraints to account for boundary effects at the ends of each partial phrase.

Hagen et al. (2004) describes more recent advances made to both acoustic and language modeling for oral-reading recognition of children's speech. Specifically, that work describes the use of cross-utterance word history modeling, position-sensitive dynamic n-gram language modeling, as well as vocal tract length normalization, speaker-adaptive training, and iterative unsupervised speaker adaptation for improved recognition. The final system was shown to have an overall word error rate of 8.0% on the speech corpus described in Section 3. This system serves as the baseline for our experiments in the remainder of the paper.

5 Event-based Word Error Analysis

While our earlier work in Hagen et al. (2003) and Hagen et al. (2004) described consistent improvements in speech recognition accuracy on children's speech, the use of raw word error rate does not reveal much information in terms of where future improvements in system performance may be obtained. Because of this, we annotated the CU Children's Read Story Corpus in terms of a set of event labels which we feel might have most relation to speech recognition error rate. Next, in Section 5.1, we describe the event labeling methodology and then provide a detailed error analysis of our baseline system in Section 5.2.

5.1 Event Labeling Methodology

The event labels for this project were chosen based on the most common types of errors children make when reading aloud. Also included in the labels are other acoustic events that occur frequently, such as breaths and pauses, which may contribute to an error made by the speech recognizer. The event labels for this study are summarized in Table 1. Common errors as stated

[3] The CU Children's Read Story Corpus is made available for research purposes (http://cslr.colorado.edu)

[4] SONIC is freely downloadable for research use from (http://cslr.colorado.edu)

[5] This corpus differs from the test corpus in Section 3.

before are word repetitions, omissions, substitutions, insertions, and self-corrections.

Although pauses (PS) are natural in speech, too many can disrupt the fluency of the read story. If a pause is extended, the recognizer may potentially insert a word (during the silence region). Similarly, we marked breath placements (BR) if they were audible. We hypothesize that words may be inserted during periods of breath if not properly accounted for by the speech recognizer. Mispronunciations (MP) tend to occur when a child is faced with word he/she is not familiar with and makes an attempt at either sounding it out (or speak fluently with an inappropriate phonetic realization). The use of wrong words (WW) is commonly a result of fast reading. The child may only read the first part of the word and guess on the rest replacing the word with one that is phonologically similar. An interjection (IJ) is any word inserted into a sentence that is not in the original text (e.g., 'um' or 'ah'). Repetitions (REP) occur when the child realizes he/she has made a mistake and self corrects him/herself usually by repeating the misread word or by beginning the sentence over again. In some cases the child catches his/her error before finishing the word and thus creating a partial word, however, since it is a conscious act by the child the word is marked as a repetition assuming he/she did repeat it to self correct.

Other important factors to be tracked by the recognizer are over-articulations (OA), hesitations (HS), non-speech segments (NS) and background noises (BN). An over-articulation is considered to be a deliberate sounding out of the word where each sound may be heard separately. A child may additionally hesitate on a word while looking ahead at the next word causing parts of the word to be elongated (e.g., stretched vowels). The non-speech sound and background noise labels are meant to indicate any noise outside of the child's reading such as a cough or a door closing. We also considered including a label for head-colds (HC), but later removed this label due to inconsistencies and subjective assessments needed.

These labels were applied to 106 read stories from the audio corpus described in Section 3. Each file was analyzed by one of three listeners and marked using these labels. Reliability between the listeners was checked by overlapping the files analyzed and comparing mark ups. The event labeling and word-level transcription of the audio corpus were conducted using the freely available *Transcriber* software[6].

Event Label & Event Description		Total Words (%)	Word Error (%)
None	No Labeled Event	92.26	5.7
REP	Word Repetition	2.46	22.4
BR	Breath	1.44	26.1
PW	Partial Word	0.70	49.6
PS	Pause	0.70	40.5
HS	Hesitation/Elongation	0.67	13.8
WW	Wrong Word	0.60	48.1
MP	Mispronunciation	0.36	36.2
BN	Background Noise	0.30	15.5
IJ	Interjection / Insertion	0.28	61.3
NS	Non-Speech Sound	0.27	58.8
OA	Over-articulation	0.10	38.3

Table 1: Event labels used in speech recognition error analysis on the CU Children's Read Story Corpus. Total words aligned to each condition are shown (in %) along with the average word error rate of the baseline system under each condition. The baseline system has a word error rate of 8.0%.

5.2 Speech Recognition Error Analysis

Using the NIST *Speech Recognition Scoring Toolkit (SCTK)*[7] we obtained the alignments of the reference word-level transcription with the hypothesized string from our baseline speech recognition system. By using the associated timing information, each word was then marked as belonging to one of the event classes shown in Table 1 (or possibly no class marking). Each word was further marked as correctly or incorrectly recognized by the speech recognizer using the scoring software. Based upon this analysis we are able to deduce the percentage of words that are output from the speech recognizer and associated with each event condition (column 2 of Table 1). We also can determine the average word-error rate for each labeled event type (column 3 of Table 1).

What is most striking from Table 1 is that the average system word error rate during non-event labeled conditions is 5.7% while the average word error rate for words associated with the labeled event conditions is 31.5%. While the speech recognizer output during the labeled events is small (approximately 7.7% of the words), the events contribute to nearly 30% of the word error rate of the system. Most troubling are instances of repeated words and breaths made by the child during read-aloud. We suggest that future progress can be made by focusing on (1) flexible n-gram language modeling which may take into account the problem of word-repetition, and (2) more accurate acoustic modeling and rejection of breath events during oral reading.

[6] http://www.etca.fr/CTA/gip/Projets/Transcriber/

[7] http://www.nist.gov/speech/tools/

6 Automatic Detection of Reading Miscues

An important aspect in an automated reading tutor is the capability of detecting reading miscues and utilizing this knowledge to provide appropriate feedback. The level of detail present in the feedback strongly depends on the event detection accuracy, which is investigated in this paper. We leave the problem of determining what feedback to provide as an area of future work. First, we define our miscue detection problem and then provide a description of the features and classifier utilized. Finally, we evaluate our miscue detection system using the baseline speech recognition system described in Section 4.

6.1 Problem Formulation

Our main criterion for detecting events in our system is based on word alignments which compare the reference transcription of the child's speech to the reference story text. Similarly to Tam et al. (2003), in order to detect reading miscues the speech recognizer's hypothesized output is aligned against the target story text using the Viterbi algorithm (i.e., *hypothesis-target* alignment). Furthermore the alignment of the human-based transcription against the story text is needed in the classification / evaluation process to determine where reading miscues actually occur (i.e., *transcription-target* alignment).

We define a reading miscue event as any instance in which the child inserts, deletes or substitutes a word during oral reading. Therefore each word spoken by the child is associated with an event label (insertion, deletion or substitution) or non-event (i.e., correct word).

Given this word-level miscue labeling of the data we can propose a detection problem. Here, each recognized word is submitted to a classifier. This classifier labels each output word as correct or incorrect (i.e., a miscue event). By thresholding the classifier output we can determine a detection rate for a given false alarm rate and therefore describe a Receiver Operating Characteristic (ROC). The detection rate is defined as the number of times the hypothesis-target and transcription-target alignments show miscues at the same position divided by the number of transcription-target miscues. The false alarm rate is defined as the number of times the hypothesis-target alignment shows a miscue at a position where the transcription-target alignment does not.

We stress that we are not interested in the exact reading miscue (wrong word, correct word but pronounced incorrectly, partial word, etc.) that occurred, which would request too specific information for a current state of the art system to give reliable feedback. Rather, we wish to design an indicator that can accurately report the detection of a miscue event whenever the text was not read correctly.

In order to be able to map one alignment to the other, the two alignments need to be synchronized. Our approach synchronizes the two alignments over the target words in the actual story text. Therefore each target word represents a unique position within both alignments. If one or more insertions occurred before a certain word in the target sentence this event is noted in a data structure attached to the specific target word stating the number of inserted words before the actual spoken word. If the word was replaced with another word in the hypothesis or transcription, the wrong word will be aligned with the actual target word, if a word is left out, no word from the hypothesis or transcription will be aligned with the specific deleted target word. Therefore the number of tokens with additional information about substitutions, deletions and insertions in the hypothesis-target alignment and transcription will be the same for both alignments and therefore word-based synchronization is ensured. To illustrate the process a short example is given. The target sentence,

it was the first day of summer vacation

might be spoken by the child (and transcribed) as,

it was ***it was*** *the* ***third*** *day of summer vacation*

and the recognition hypothesis might state,

it ***it*** *was the first day vacation*

Therefore this transcription would have two insertions and one substitution events. The hypothesis would have one insertion and two deletions ("of summer"). The alignments along with the attached information are shown in Table 2. The *miscue* columns indicate an event occurring at a specific position in the target text or right before it in the case of one or more insertions before a certain word.

Story (Target)	Trans. (Ref.)	Actual Miscue	Recognizer (Hyp.)	Hyp. Miscue
it	it	0-0-0	it	0-0-0
was	was	0-0-0	was	0-0-1
the	the	0-0-2	the	0-0-0
first	third	1-0-0	first	0-0-0
day	day	0-0-0	day	0-0-0
of	of	0-0-0	<no_word>	0-1-0
summer	summer	0-0-0	<no_word>	0-1-0
vacation	vacation	0-0-0	vacation	0-0-0

Table 2: Transcription-target alignment and hypothesis-target alignment with substitution-deletion-insertion (s-d-i) miscue annotation.

This setup enables us to compute the detection and false alarm rates based on the synchronized alignments. Within the Viterbi-alignment process a soft decision is made whether to classify a word as a substitution or not. If the phonemes of the hypothesized word match the phonemes of the target word by at least 75% (determined by phoneme alignment) the word is accepted as correct. This softer decision overcomes less important events like misses of an 's/z' sound at the end of a word (e.g., "piano" vs. "pianos").

6.2 Features for Miscue Detection

The alignment based miscue detection is only capable of providing a single operating point (detection rate / false alarm rate). We next introduce additional features which allow us to threshold the classifier output and allow the system to operate at any point along the ROC curve.

In order to be able to operate the detector at different levels of sensitivity additional features to the alignment used in a classifier are a useful extension. The features we chose are,

- the word alignment
  (either 1 if the hypothesized word aligns to the target story word or 0 otherwise)
- the speech recognizer language model score
  (computed per word)
- the speech recognizer acoustic score
  (per word, normalized by frame count)
- the length of the pause in seconds before the current word (0 if no pause exists)
- the number of phonemes in the current word

The alignment is obtained as discussed in Section 6.1. The language model score and the normalized acoustic score are indicators for the quality of the match between the hypothesized word's model and the observed features. The length of the pause before a word indicates a hesitation that might be a hint for a reading irregularity. The number of phonemes should reflect the assumption that longer words are generally harder to read, especially for bad readers.

6.3 Classifier Formulation

We trained a linear classifier based on the features discussed above. The use of a linear classifier was motivated by earlier work of Hazen et al. (2001) which demonstrated that such a classifier can generate acceptable performance for speech recognizer confidence estimation given that the decision surface is relatively simple. The classifier can be expressed as,

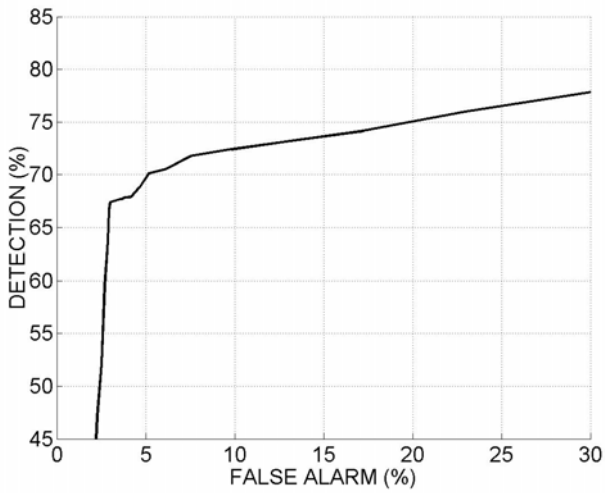
$$r = \vec{p}^{\,T} \vec{f}$$

where $\vec{p}$ is the trained classification vector and $\vec{f}$ is the feature vector described above. The final classification is based on a threshold value. If r is greater than the threshold value, the instance under investigation is classified as a miscue, otherwise as a non-event. By varying r over a certain range the receiver operating characteristic (ROC) curve can be obtained.

6.4 Evaluation

The data set used to train the classifier consists of 50% of the CU Children's Read Story Corpus randomly chosen such that age and grade levels are distributed similarly to the entire corpus. The training examples represent both miscue and non-miscue events. The miscues are those examples that represent substitutions, deletions, or insertions within the transcription-target alignment. The negative examples are chosen from the non-miscue examples. There are 4,875 miscue and 8,715 non-miscue examples used to train the classifier.

We tested the classifier on the remaining 50% of the corpus. There are approximately 5,000 miscues in the test set. The ROC curve resulting from the classification system applied to the test set is shown in Figure 2. It can be seen that the overall performance has a relatively high detection rate of 67% with a false alarm rate of less than 3.0%. With the detection rate adjusted to 70% and higher the false alarm rate increases rapidly.



DT (%)	55.0	60.0	65.0	70.0	75.0	80.0
FA (%)	2.6	2.7	2.9	5.1	19.8	36.4

Figure 2: Detection rate vs. false alarm rate ROC for the CU Children's Read Story corpus. Example operating points are shown below.

7 Conclusions

In this paper we have described the Colorado Literacy Tutor (CLT) which aims to improve literacy and student achievement in public schools. We extended on our previous work in several novel aspects. First, we have collected and annotated a children's speech corpus in terms of a set of labeled event conditions which we believe strongly correlate to speech recognition error. In fact while these events make up nearly 8% of the data, they were shown to account for approximately 30% of the word errors in a state-of-the-art speech recognition system. To our knowledge, previous work has not considered such a detailed word error analysis on a children's speech corpus. We then provided our initial framework for detecting oral reading miscues. Using a simple linear classifier and using features derived from a speech recognizer, we demonstrated that 67% of reading miscues can be detected at a false alarm rate of 3%. While this system appears to outperform the previous results presented in Tam et al. (2003), we point out that there is currently no standardized test set available to directly compare those systems. Therefore, the audio corpus and event labeling presented in this paper will be made available to researchers to promote community-wide benchmarking. In the future we plan to correlate the miscue detection performance with the event labeling strategy outlined in Section 5 of the paper. We expect that such an error analysis will continue to provide insight to areas for system development.

8 Acknowledgements

This work was supported by grants from the National Science Foundation's ITR and IERI Programs under grants NSF/ITR: REC-0115419, NSF/IERI: EIA-0121201, NSF/ITR: IIS-0086107, NSF/IERI: 1R01HD-44276.01; and the Coleman Institute for Cognitive Disabilities. The views expressed in this paper do not necessarily represent the views of the NSF.

References

R. Cole, S. van Vuuren, B. Pellom, K. Hacioglu, J. Ma, J. Movellan, S. Schwartz, D. Wade-Stein, W. Ward, J. Yan. 2003. *Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human Computer Interaction*. Proceedings of the IEEE, Vol. 91, No. 9, pp. 1391-1405.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. 1990. *Indexing by Latent Semantic Analysis*. Journal of the Society for Information Science, vol. 41, no. 6, pp. 391-407.

J. Fogarty, L. Dabbish, D. Steck, and J. Mostow. 2001. *Mining a Database of Reading Mistakes: For What should an Automated Reading Tutor Listen?* In J. D. Moore, C. L. Redfield, and W. L. Johnson (Eds.), Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future, pp. 422-433.

A. Hagen, B. Pellom, and R. Cole. 2003. *Children's Speech Recognition with Application to Interactive Books and Tutors*. ASRU-2003, St. Thomas, USA.

A. Hagen, B. Pellom, S. Van Vuuren, R. Cole. 2004. *Advances in Children's Speech Recognition within an Interactive Literacy Tutor*. HLT-NAACL, Boston Massachusetts, USA.

T. Hazen, S. Seneff, and J. Polifroni. 2002. *Recognition Confidence Scoring and its Use in Speech Understanding Systems*. Computer Speech and Language, Vol. 16,No. 1, pp. 49-67.

E. Kintsch, D. Steinhart, G. Stahl, C. Matthews, R. Lamb, and LRG. 2000. *Developing Summarization Skills through the Use of LSA-based Feedback*. Interactive Learning Environments, Vol. 8, pp. 87-109.

T. Landauer and S. Dumais. 1997. *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge*. Psych. Review, Vol. 104, pp. 211-240.

M. McCandless. 1992. *Word Rejection for a Literacy Tutor.* Bachelor of Science Thesis, MIT.

J. Mostow, G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, M. B. Sklar, and B. Tobin. 2003. *Evaluation of an Automated Reading Tutor that Listens: Comparison to Human Tutoring and Classroom Instruction*. Journal of Educational Computing Research, 29(1), 61-117.

J. Mostow, J. Beck, S. Winter, S. Wang, and B. Tobin. 2002. *Predicting Oral Reading Miscues*. ICSLP-02, Denver, Colorado.

J. Mostow, S. Roth, A. G. Hauptmann, and M. Kane. 1994. *A Prototype Reading Coach that Listens*. AAAI-94, Seattle, WA, pp. 785-792.

B. Pellom. 2001. *SONIC: The University of Colorado Continuous Speech Recognizer*. Technical Report TR-CSLR-2001-01, University of Colorado.

B. Pellom, K. Hacioglu. 2003. *Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task*. Proc. ICASSP, Hong Kong.

K. Shobaki, J.-P. Hosom, and R. Cole. 2000. *The OGI Kids' Speech Corpus and Recognizers*. Proc. ICSLP-2000, Beijing, China.

D. Steinhart. 2001. *Summary Street: An Intelligent Tutoring System for Improving Student Writing through the Use of Latent Semantic Analysis*. Ph.D. Dissertation, Dept. Psychology, Univ. of Colorado, Boulder, CO.

Y-C. Tam, J. Mostow, J. Beck, and S. Banerjee. 2003. *Training a Confidence Measure for a Reading Tutor that Listens*. Proc. Eurospeech, Geneva, Switzerland, 3161-3164.

U. Yapanel, J. H.L. Hansen. 2003. *A New Perspective on Feature Extraction for Robust In-vehicle Speech Recognition*. Proc. Eurospeech, Geneva, Switzerland.

V. Zue, S. Seneff, J. Polifroni, H. Meng, J. Glass. 1996. *Multilingual Human-Computer Interactions: From Information Access to Language Learning*. ICSLP-96, Philadelphia, PA.