Advances in Children's Speech Recognition within an Interactive Literacy Tutor

Andreas Hagen, Bryan Pellom, Sarel Van Vuuren, and Ronald Cole
Center for Spoken Language Research
University of Colorado at Boulder
http://cslr.colorado.edu

Abstract[1]

In this paper we present recent advances in acoustic and language modeling that improve recognition performance when children read out loud within digital books. First we extend previous work by incorporating cross-utterance word history information and dynamic n-gram language modeling. By additionally incorporating Vocal Tract Length Normalization (VTLN), Speaker-Adaptive Training (SAT) and iterative unsupervised structural maximum a posteriori linear regression (SMAPLR) adaptation we demonstrate a 54% reduction in word error rate. Next, we show how data from children's read-aloud sessions can be utilized to improve accuracy in a spontaneous story summarization task. An error reduction of 15% over previous published results is shown. Finally we describe a novel real-time implementation of our research system that incorporates time-adaptive acoustic and language modeling.

1 Introduction

Pioneering research by MIT and CMU as well as more recent work by the IBM *Watch-me-Read* Project have demonstrated that speech recognition can play an effective role in systems designed to improve children's reading abilities (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). Upon detecting reading miscues, the tutor provides appropriate feedback to the child. Recent results show that such automated reading tutors can improve student achievement (Mostow et al, 2003). Providing real time feedback by highlighting words as the are read out loud is the basis of at least one commercial product today (http://www.soliloquy.com).

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 speech recognition, facial animation, and natural language understanding capabilities to teach children to read and comprehend text. Interactive Books are designed to improve student achievement by helping students to learn to read fluently, to acquire new knowledge through deep understanding of what they read, to make connections to other knowledge, and to express their ideas concisely through spoken or written summaries. Transcribed spoken summaries can be graded automatically to provide feedback to the student about their comprehension.

During reading out loud activities in Interactive Books, the goal is to design a computer interface and speech recognizer that combine to teach the student to read fluently and naturally. Here, speech recognition is used to track a child's position within the text during read-aloud sessions in addition to providing timing and confidence information which can be used for reading assessment. The speech recognizer must follow the students verbal behaviors accurately and quickly, so the cursor (or highlighted word) appears at the right place and right time when the student is reading fluently, and pauses when the student hesitates to sound out a word. The recognizer must also score mispronounced words accurately so that the student can revisit these words and receive feedback about their pronunciation after completing a paragraph or page (since highlighting hypothesized mispronounced words when reading out loud may disrupt fluent reading behavior).

In this paper we focus on the problem of speech recognition to track and provide feedback during reading out loud and to transcribe spoken summaries of text. Specifically, we describe several new methods for in-

[1] This work was supported in part by grants from the National Science Foundation's Information Technology Research (ITR) Program and the Interagency Educational Research Initiative (IERI) under grants NSF/ITR: REC-0115419, NSF/IERI: EIA-0121201, NSF/ITR: IIS-0086107, NSF/IERI: 1R01HD-44276.01, NSF: INT-0206207; and the Coleman Institute for Cognitive Disabilities. The views expressed in this paper do not necessarily represent the views of the NSF.

corporating language modeling knowledge into the read aloud task. In addition, through use of speaker adaptation, we also demonstrate the potential for significant gains in recognition accuracy. Finally, we leverage improvements in speech recognition for read aloud tracking to improve performance for spoken story summarization. Work reported here extends previous work in several important ways: by integrating the research advances into a real time system, and by including time-adaptive language modeling and time-adaptive acoustic modeling of the child's voice into the system.

The paper is organized as follows. Sect. 2 describes our baseline speech recognition system and reading tracking method. Sect. 3 presents our rationale for using word-error-rate as a measure of performance. Sect. 4 describes the read aloud and story summarization corpora used in this work. Sect. 5 describes and evaluates proposed improvements in a read aloud speech recognition task. Sect. 6 describes how these improvements translate to improved recognition of story summaries produced by a child. Sect. 7 details our real-time system implementation.

2 Baseline System

For this work we use the SONIC speech recognition system (Pellom, 2001; Pellom and Hacioglu, 2003). 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 HMMs. For children's speech, the recognizer has been trained on 46 hours of data from children in grades K through 9 extracted from the CU Read and Prompted speech corpus (Hagen et al., 2003) and the OGI Kids' speech corpus (Shobaki et al., 2000). Further, the baseline system utilizes PMVDR cepstral coefficients (Yapanel and Hansen, 2003) for improved noise robustness.

During read-aloud operation, 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 book 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.

3 Evaluation Methodology

There are many different ways in which speech recognition can be used to serve children. In computer-based literacy tutors, speech recognition can be used to measure children's ability to read fluently and pronounce words while reading out loud, to engage in spoken dialogues with an animated agent to assess and train comprehension, or to transcribe spoken summaries of stories that can be graded automatically. Because of the variety of ways of using speech recognition systems, it is critically important to establish common metrics that are used by the research community so that progress can be measured both within and across systems.

For this reason, we argue that word error rate calculations using the widely accepted NIST scoring software provides the most widely accepted, easy to use and highly valid metric. In this scoring procedure, word error rate is computed strictly by comparing the speech recognizer output against a known human transcription (or the text in a book). Of course, authors are free to define and report other measures, such as detection/false alarm curves for useful events such as reading miscues. However, such analyses should always supplement reports of word error rates using a single standardized measure. Adopting this strategy enables fair and balanced comparisons within and across systems for any speech data given a known word-level transcription.

4 Experimental Data

For all experiments in this paper we use speech data and associated transcriptions from 106 children (grade 3: 17 speakers, grade 4: 28 speakers, and grade 5: 61 speakers) who were asked to read one of ten stories and to provide a spoken story summary. The 16 kHz audio data contains an average of 1054 words (min 532 words; max 1926 words) with an average of 413 unique words per story. The resulting summaries spoken by children contain an average of 168 words.

5 Improved Read-Aloud Recognition

Baseline: Our baseline read-aloud system utilizes a trigram language model constructed from a normalized version of the story text. Text normalization consists primarily of punctuation removal and determination of sentence-like units. For example,

It was the first day of summer vacation. Sue and Billy were eating breakfast. "What can we do today?" Billy asked.

is normalized as:

<s> IT WAS THE FIRST DAY OF SUMMERVACATION</s>
<s> SUE AND BILLY WERE EATING BREAKFAST</s>
<s> WHAT CAN WE DO TODAY </s>
<s> BILLY ASKED </s>

The resulting text is used to estimate a back-off trigram language model. We stress that only the story text is used to construct the language model. Details on the story texts are provided in Hagen et al. (2003). Note that the sentence markers (<s> and </s>) are used to represent positions of expected speaker pause. This baseline system is shown in Table 1(A) to produce a 17.4% word error rate.

Improved Sentence Context Modeling: It is important in the context of this research to note that children do not pause between each estimated sentence boundary. Instead, many children read fluently across phrases and sentences, where more experienced readers would pause. For this reason, we improved upon our baseline system by estimating language model parameters using a combined text material that is generated both with and without the contextual sentence markers (<s> and </s>). Results of this modification are shown in Table 1(B) and show a reduction in error from 17.4% to 13.5%.

Improved Word History Modeling: Most speech recognition systems operate on the utterance as a primary unit of recognition. Word history information typically is not maintained across segmented utterances. However, in our text example, the words “do today” should provide useful information to the recognizer that “Billy asked” may follow. We therefore modify the recognizer to incorporate knowledge of previous utterance word history. During token-passing search, the initial word-history tokens are modified to account for the fact that the incoming sentence may be either the beginning of a new sentence or a direct extension of the previous utterance’s word-end history. Incorporating this constraint lowers the word error rate from 13.5% to 12.7% as shown in Table 1(C).

Dynamic n-gram Language Modeling: During story reading we can anticipate words that are likely to be spoken next based upon the words in the text that are currently being read aloud. To account for this knowledge, we estimate a series of position-sensitive n-gram language models by partitioning the story into overlapping regions containing at most 150 words (i.e., each region is centered on 50 words of text with 50 words before and 50 words after). For each partition, we construct an n-gram language model by using the entire normalized story text in addition to a 10x weighting of text within the partition. Each position-sensitive language model therefore contains the entire story vocabulary. We also compute a general language model estimated solely from the entire story text (similar to Table 1(C)). At run-time, the recognizer implements a word-history buffer containing the most recent 15 recognized words. After decoding each utterance, the probability of the text within the word history buffer is computed using each of the position-sensitive language models. The language model with the highest probability is selected for the first-pass decoding of the subsequent utterance. This modification decreases the word error rate from 12.7% to 10.7% (Table 1(D)).

Vocal Tract Normalization and Acoustic Adaptation: We further extend on our baseline system by incorporating the Vocal Tract Length Normalization (VTLN) method described in Welling et al. (1999). Based on results shown in Table 1(E), we see that VTLN provides only a marginal gain (0.1% absolute). Our final set of acoustic models for the read aloud task are both VTLN normalized and estimated using Speaker Adaptive Training (SAT). The SAT models are determined by estimating a single linear feature space transform for each training speaker (Gales, 1997). The means and variances of the VTLN/SAT models are then iteratively adapted using the SMAPLR algorithm (Siohan, 2002) to yield a final recognition error rate of 8.0% absolute (Table 1(G)). By combining all of these techniques, we achieved a 54% reduction in word error rate relative to the baseline system.

Experimental Configuration		Word Error Rate (%)	
		MFCC	PMVDR
(A)	Baseline: single n-gram language model	17.7%	17.4%
(B)	(A) + Begin/End Sentence Context Modeling	14.0%	13.5%
(C)	(B) + between utterance word history modeling	13.0%	12.7%
(D)	(C) + dynamic n-gram language model	11.0%	10.7%
(E)	(D) + VTLN	10.9%	10.6%
(F)	(E) + VTLN/SAT + SMAPLR (iteration 1)	8.2%	8.2%
(G)	(E) + VTLN/SAT + SMAPLR (iteration 2)	8.0%	8.0%

Table 1: Recognition of children’s read out-loud data.

6 Improved Story Summary Recognition

One of the unique and powerful features of our interactive books is the notion of assessing and training comprehension by providing feedback to the student about a typed summary of text that the student has just read (Cole et al., 2003). Verbal input is especially important for younger children who often can not type well. Utilizing summaries from the children’s speech corpus, Hagen et al. (2003) showed that an error rate of 42.6% could be achieved. The previous work, however, did not consider utilizing the read story material to provide improved initial acoustic models for the summarization task. In Table 2 we demonstrate several findings using a language model trained on story text and example summaries produced by children (leaving out data from the child under test). Without any adaptation the error rate is 47.1%. However, utilizing adapted models from the read stories (see Table 1(G)) provides an initial performance gain of nearly 10% absolute. Further use SMAPLR adaptation reduces the error rate to 36.1%.

Experimental Configuration		Word Error Rate (%)	
		MFCC	PMVDR
(A)	Baseline / no adaptation	47.0%	47.1%
(B)	Read-aloud adapted models (VTLN/SAT)	37.2%	38.0%
(C)	(B) + SMAPLR adaptation iteration #1	36.0%	36.6%
(D)	(C) + SMAPLR adaptation iteration #2	35.1%	36.1%

Table 2: Recognition of spontaneous story summaries

7 Practical Real-Time Implementation

The research systems described in Sect. 5 and 6 do not operate in real-time since multiple adaptation passes over the data are required. To address this issue, we have implemented a real-time system that operates on small pipelined audio segments (250ms on average). When evaluated on the read-aloud task (Sect. 5), the initial baseline system achieves an error rate of 19.5%. This system has a real-time factor of 0.56 on a 2.4 GHz Intel Pentium 4 PC with 512MB of RAM. When integrated, the proposed methods show the error rate can be reduced from 19.5% to 12.7% (compare with 10.7% error research system in Table 1(D)). The revised system which incorporates dynamic language modeling operates 35% faster than the single language model method while also reducing the variance in real-time factor for each processed chunk of audio. Further gains are possible by incorporating adaptation in an incremental manner. For example, in Table 3(C) a real-time system that incorporates incremental unsupervised maximum likelihood linear regression (MLLR) adaptation of the Gaussian means is shown. *This final real-time system simultaneously adapts both language and acoustic model parameters during system use*. The system is now being refined for deployment in classrooms within the CLT project. We were able to further improve the system after the submission deadline. The current WER on the story read aloud task improved to 7.6%; while a WER of 32.2% was achieved on the summary recognition task. The improvements are due to the inclusion of a breath model and the additional use of audio data from 103 second graders for more accurate acoustic modeling.

System Description		PMVDR Front-End	
		WER (%)	RTF
(A)	Baseline: single LM	19.5%	0.56 $(\sigma^2=0.11)$
(B)	Proposed System	12.7%	0.36 $(\sigma^2=0.06)$
(C)	(B) + Incremental MLLR adaptation	11.5%	0.80 $(\sigma^2=0.33)$

Table 3: Evaluation of real-time read out-loud system.

References

V. Zue, S. Seneff, J. Polifroni, H. Meng, J. Glass (1996). “Multilingual Human-Computer Interactions: From Information Acess to Language Learning,” ICSLP-96, Philadelphia, PA

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

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

J. Mostow, J. Beck, S. Winter, S. Wang, and B. Tobin (2002). “Predicting oral reading miscues” ICSLP-02, Denver, Colorado.

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

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

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

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

B. Pellom and K. Hacioglu (2003). "Recent Improvements in the CU Sonic ASR System for Noisy Speech: The SPINE Task", ICASSP-2003, Hong Kong, China.

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

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

L. Welling, S. Kanthak, and H. Ney. (1999) "Improved Methods for Vocal Tract Length Normalization", ICASSP, Phoenix, Arizona.

M. Gales (1997). Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition", Tech. Report, CUED/F-INFENG/TR291, Cambridge University.

O. Siohan, T. Myrvoll, and C.-H. Lee (2002) "Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation", Computer, Speech and Language, 16, pp. 5-24.