

# Highly accurate children's speech recognition for interactive reading tutors using subword units

Andreas Hagen, Bryan Pellom \*, Ronald Cole

*Center for Spoken Language Research, University of Colorado at Boulder, 1777 Exposition Drive, Suite #171, Boulder, CO 80301, USA*

Received 15 December 2005; received in revised form 20 February 2007; accepted 9 May 2007

## Abstract

Speech technology offers great promise in the field of automated literacy and reading tutors for children. In such applications speech recognition can be used to track the reading position of the child, detect oral reading miscues, assessing comprehension of the text being read by estimating if the prosodic structure of the speech is appropriate to the discourse structure of the story, or by engaging the child in interactive dialogs to assess and train comprehension. Despite such promises, speech recognition systems exhibit higher error rates for children due to variabilities in vocal tract length, formant frequency, pronunciation, and grammar. In the context of recognizing speech while children are reading out loud, these problems are compounded by speech production behaviors affected by difficulties in recognizing printed words that cause pauses, repeated syllables and other phenomena. To overcome these challenges, we present advances in speech recognition that improve accuracy and modeling capability in the context of an interactive literacy tutor for children. Specifically, this paper focuses on a novel set of speech recognition techniques which can be applied to improve oral reading recognition. First, we demonstrate that speech recognition error rates for interactive read aloud can be reduced by more than 50% through a combination of advances in both statistical language and acoustic modeling. Next, we propose extending our baseline system by introducing a novel token-passing search architecture targeting subword unit based speech recognition. The proposed subword unit based speech recognition framework is shown to provide equivalent accuracy to a whole-word based speech recognizer while enabling detection of oral reading events and finer grained speech analysis during recognition. The efficacy of the approach is demonstrated using data collected from children in grades 3–5, namely 34.6% of partial words with reasonable evidence in the speech signal are detected at a low false alarm rate of 0.5%.

© 2007 Elsevier B.V. All rights reserved.

*Keywords:* Literacy tutors; Subword unit based speech recognition; Language modeling; Reading tracking

## 1. Introduction

In recent years, automated reading tutors that utilize speech recognition technology to track and assess a child's reading ability have become more feasible due to increased computer power and advances in accurate and efficient methods for speech recognition (Mostow et al., 1994; Cole et al., 2003). Previous studies have considered acoustic analysis of children's speech (Lee et al., 1997; Lee et al.,

1999; Li and Russell, 2002). This work has shed light onto the challenges faced by systems that will be developed to automatically recognize and effectively model children's speech patterns. For example, it has been shown that children below the age of 10 exhibit a wider range of vowel durations relative to older children and adults, larger spectral and suprasegmental variations, and wider variability in formant locations and fundamental frequencies in the speech signal. In recent years, several studies have attempted to address these issues by adapting the acoustic features of children's speech to match that of acoustic models trained from adult speech (Potamianos et al., 1997; Das et al., 1998; Potamianos and Narayanan, 2003;

\* Corresponding author. Tel.: +1 303 735 5382; fax: +1 303 735 5072.  
E-mail addresses: [andreash@cslr.colorado.edu](mailto:andreash@cslr.colorado.edu) (A. Hagen), [pellom@cslr.colorado.edu](mailto:pellom@cslr.colorado.edu) (B. Pellom), [cole@cslr.colorado.edu](mailto:cole@cslr.colorado.edu) (R. Cole).

Giuliani and Gerosa, 2003; Gustafson and Sjolander, 2002). Approaches of this sort have included vocal tract length normalization as well as spectral normalization. Each of these earlier studies point to lack of children's acoustic data needed to estimate speech recognition parameters relative to the over abundance of existing resources for adult speech recognition. More recently, corpora for children's speech recognition have begun to emerge. In (Eskenazi et al., 1996) a small corpus of children's speech was collected for use in interactive reading tutors and led to a complete children's speech recognition system (Aist et al., 1998). In (Shobaki et al., 2000), a more extensive corpus consisting of 1100 children in grades K through 10 was collected in Oregon for US English and used to develop a speech recognition system for isolated word and finite-state grammar vocabularies. The development and increasing availability of speech and language resources for has resulted in the development of several reading tutors which support some degree of analysis using speech recognition. Such systems claim to listen to children effectively while providing valuable feedback (Mostow et al., 1994).

In 2003 we developed a baseline reading recognition system based on the SONIC speech recognizer (Pellom, 2001; Pellom and Hacıoglu, 2003), that was trained on 50 h of children's speech (grade K through 5) and used a trigram language model trained on the story text. This system had a Word Error Rate (WER) of 16.5% on the test set described below in Section 3.2. Analysis of data of children reading stories out loud has yielded several insights. For example, an analysis based on the speech transcripts and the actual story texts showed that children who are early readers often do not pause at expected points of punctuation. Early readers were found to pause at wrong positions (where there were no punctuation marks like commas, periods, question marks, etc. that would indicate a pause for experienced readers) about 16 times in an average story of 1054 words. In fact, for our test set which is presented in Section 3.2, the 106 speakers together ignored 2379 of 7929 pauses indicated by punctuations in the text; therefore 30% of all punctuation marks were not realized as pauses during oral reading. It is interesting to look at the relative number of pauses missed by grade level. Third graders ignored 24.5% of all pauses, fourth graders 27.8%, and fifth graders ignored 31.2%. So at a higher grade children tend to read over punctuation marks more often. At least a 12% relative increase in pause misses could be observed per grade level transition from third to fifth graders. This is not surprising since children at higher grade levels are reading at faster rates on average. In order to exploit these phenomena we developed new techniques, extending the baseline system, which are presented in Section 4. The system could be improved significantly by more than 50% in WER.

In (Lee et al., 2004), we examined the types of speech recognition errors made by the SONIC system during recognition of oral readings by children. The corpus was labeled by hand by three annotators into a set of event conditions (e.g., word repetition, mispronunciation, sounding

out of words, etc.). It was found that while 8% of the labeled corpus contained event conditions (reading miscues of one sort or another), the events themselves described almost 30% of the word errors made by the speech recognizer. This research informed the need for a subword unit approach to children's speech recognition in order to model the types of errors that occur during reading out loud. Mispronunciations and partial words, which account for approximately 34% of reading miscues, can be described on the subword unit level. There is thus a mismatch between the need to model dysfluencies as children read aloud in terms of subword units and design of current state of the art large vocabulary speech recognition systems. Most state of the art speech recognition systems compactly and efficiently represent the search space by representing the word lexicon in the form of a prefix tree, and thus do not easily allow for representation and recognition of oral dysfluencies that occur during children's reading out loud. Therefore in this work we propose and implement a new technique for modeling words as constituent parts or subword units. Subword units will be shown to enable the detection of events occurring on the subword level, as for example partial words.

The paper is organized as follows. We provide detailed information on related work in Section 2. Section 3 introduces the speech corpora used for experiments. Section 4 describes advanced language modeling techniques for reading tutors and Section 5 describes our hybrid word/subword unit recognition system. By combining several techniques we show a 52% relative improvement in word error rates during oral reading and the proposed subword unit recognition system additionally enables special event analysis such as partial word detection.

## 2. Related work

In the following sections we summarize previous work related to interactive literacy tutors which utilize children's speech recognition.

### 2.1. Literacy tutors

#### 2.1.1. MIT's literacy tutor

Earlier work at the Spoken Language Systems group at MIT's Laboratory for Computer Science resulted in the development of an automated literacy tutor (McCandless, 1992). In this system, text was presented on a screen for the user to read. The system listened to what the user spoke using speech recognition and then automatically decided if a word was read accurately. In cases of mispronounced or poorly articulated words, the system provided interactive feedback.

The author concentrated on the tutor's algorithm designed to accept or reject words while many other important issues regarding the interface and the real-time behavior was left for future work. The speech recognizer used in the MIT literacy tutor was the SUMMIT speech recogni-

tion system. It was based on a Viterbi search through the lexical network. The author defined a rejection algorithm, which assumed a speech waveform and a word transcription as input, and generated a variable length vector of word transition times as output. Various methods for rejection were presented in the work and were evaluated based on the false acceptance and false rejection rate. At a 5% false rejection rate, the MIT system provided a 7.3% false acceptance rate.

### 2.1.2. CMU's Project LISTEN reading tutor

At Carnegie Mellon University (CMU), Mostow and colleagues have developed a reading coach based on the Sphinx-II continuous speech recognition (Mostow et al., 1994). In Project LISTEN, the speech recognizer is used to listen to children read connected text, detect oral reading miscues, and automatically trigger pedagogically appropriate interventions. The concept the reading coach follows is called "shared reading". In this paradigm, the tutor implements a combination of reading and listening. The tutor intervenes when the reader misreads one or more words in the sentence, gets stuck, or clicks on a word to get help. Interventions assist word identification and the attentional bottleneck. The tutor waits to intervene until the end of a sentence unless the reader gets stuck or clicks for help. This has pedagogical reasons (does not disrupt flow) and technical reasons (system is too slow, cannot interrupt gracefully).

Project LISTEN's reading tutor displays one sentence at a time to enable smooth interventions and to improve recognition accuracy due to more precise language model bigram estimation. Early results presented in (Mostow et al., 1994) showed that the assisted reading level came out to be 0.6 years higher than independent reading level on a widely used test of oral reading (Spache, 1981). Within the reading coach speech recognition plays a critical role in deciding when the speaker deviates from the text. In cases of deviation interventions might be necessary. According to the authors these interventions require the following speech analysis capabilities: detect misread words, detect when the reader reaches the end of a sentence, and detect when the reader gets stuck. Misread words can be detected by alignment of the hypothesis with the actual text; the end of sentence detection is done by comparison of the hypothesized words with the end of sentence word. A detection mechanism when a reader gets stuck can be implemented by time limit. The speech recognizer described in Mostow's initial 1994 reading tutor used adult acoustic models and a bigram model trained on the current sentence's words, and additionally modeling word truncations, jumps, and repetitions.

Various modifications to the basic reading coach have been performed since 1994. Experiments that are most relevant to our own work are the training of a confidence measure for the correctness of a certain recognized word (Tam et al., 2003) and the ideas to sharpen the language modeling used in the tutor (Banerjee et al., 2003a). Work

regarding the prediction of oral reading miscues (Fogarty et al., 2001; Mostow et al., 2002; Banerjee et al., 2003b) is also of interest. The purpose of the confidence measure is to estimate the probability that a given word in a sentence was read correctly.

### 2.1.3. University of Colorado's Foundations to Literacy program

*Foundations to Literacy* is a comprehensive, scientifically-based reading program designed for beginning and poor readers (Cole et al., 2003, 2006; Wise et al., 2005). It consists of a set of tightly integrated computer-based multi-modal learning applications in which children interact with a Virtual Tutor, Marni, to learn to read well and to comprehend stories. It is designed to combine the benefits of scientifically-based reading research, individualized face-to-face instruction with a virtual tutor, and immersion in multi-modal computer interactions. It teaches foundational reading skills (alphabet, phonological awareness, decoding) to children who have difficulty with word reading, and it also supports and improves listening and reading comprehension and fluency after these students have mastered accurate word reading skills.

Within the *Foundations to Literacy* program, Interactive Books teach students to read fluently and comprehend text. They enable a wide range of user behaviors and interactions with the Virtual Tutor, including: (a) having the story, or any portion of it, narrated by the animated character with accurate visible speech, (b) enabling the student to click on individual words or sentences while reading silently or out loud to have them spoken by the agent, (c) providing feedback to the student (using automatic speech recognition) while reading aloud by having a cursor follow the students as they are producing words, and (d) having the student respond to questions posed by the agent (by clicking on objects in images or answering multiple choice questions).

Since 2003, *Foundations to Literacy* has been deployed in about 50 kindergarten, first grade and second grade second classrooms in Colorado schools, and has produced significant learning gains in letter and word recognition skills. From about the middle of first grade, students using the program read out loud within Interactive Books, with real-time feedback provided by the SONIC speech recognizer. Improvements to the SONIC speech recognition system, based on the system described below, have been implemented into Interactive Books within a reading out loud tracker designed by (van Vuuren et al., 2006).

## 2.2. Children's speech recognition and analysis

Lee et al. (1999) conducted an extensive analysis on properties of children's speech between the ages 5 and 18. The study showed interesting results about vowel duration, duration of the phonemes, sentence durations, and variability associated with formants and spectral envelopes.

Statistics were gathered to see how these variables change with age.

Vowel durations are significantly longer for 5 and 6 year old children. Between ages 10 and 15, vowel durations decrease significantly; beyond age 15, no further significant changes in average duration change were found. The effect of gender on vowel duration is not significant. Also variability of vowel durations, both between and within subjects, is significantly higher at a younger age. Duration and within-subject variability reach adult level at about age 12. Fricatives show similar properties, with the durations of fricatives longer for 5 and 6 year old children, with durations decreasing significantly between the ages of 10–12 and reaching a stable duration at age 13. Interestingly, sentence duration decreases almost linearly between from age 7 to 14. The authors suspect that the drop in duration between age 7 and 14 might be due to reading ability as well as pause durations, besides the obvious speaking rate. The fundamental frequency for male speakers drops significantly from ages 11 to 13 and between ages 12 to 15, due to pubertal pitch change. For female speakers the pitch drop between age 7 and 12 is significant. For intra-subject pitch variability the effect of age is again significant, younger children have higher variability. Average cepstrum distances between two repetitions of the same vowel show higher variability for younger children. This variability decreases to adult level at about age 14. The change after age 11 is not significant. The sum total of these variabilities indicates why speech recognition was found to be relatively inaccurate for young children.

Li and Russell (2002) investigated the first three formants beyond the fundamental frequency for different vowels in children's speech and the effect of reduced bandwidth on speech recognition accuracy. It was found that the average value for F1 is 182.5 Hz higher for children compared to adults. The differences in F2 and F3 are 669.3 Hz and 1008.5 Hz, respectively. F3 for children sometimes exceeds 4 kHz and therefore recognition of children's speech under reduced bandwidth, i.e., telephone speech, is affected negatively. This effect was demonstrated by comparing recognition results for adult and children's speech. The children's speech recognition accuracy decreased earlier than the adults' when the bandwidth was gradually reduced from 16 kHz to 2 kHz, and the accuracy loss was more substantial. For children the error rate began to increase at 6 kHz and significantly increases between 6 kHz and 4 kHz, for adult speech the error rate increase is insignificant until 4 kHz.

Arcy et al. (2004) introduce two new speech corpora, one with over 14 h of read speech from 159 British English speaking children, the second one a corpus with emotional speech of 1 h 23 min from 30 children. Among other experiments the authors tested if age-dependent models are beneficial over one acoustic model trained from all data available. This could not be confirmed and is therefore consistent with our earlier results in (Hagen et al., 2003), where it was shown that one acoustic model trained from

all available data using vocal tract length normalization outperforms various age-dependent models.

Cosi and Pellom (2005) investigated speech recognition performance for Italian speech using ITC-irst Children's Corpus (Giuliani and Gerosa, 2003). The authors showed that recognition accuracy of children's speech is lower when systems use adult acoustic models relative to children's acoustic models. Even with acoustic adaptation this gap remains at about 20% relative. This result shows the necessity for children's speech databases that are essential for accurate recognition of children's speech given current speech technology.

### 3. CU Children's audio speech corpora

The research conducted in this paper makes use of two new children's speech corpora collected by (Cole and Pellom, 2006a,b) at the University of Colorado.

#### 3.1. University of Colorado Prompted and Read Children's speech corpus

The CSLR Prompted and Read Children's speech corpus consists of transcribed speech data collected from 663 Kindergarten through fifth grade children producing isolated words, sentences, and short spontaneous stories. The protocol is described in (Cole and Pellom, 2006a). Table 1 provides the number of speakers per grade level.

Each speaker produced approximately 100 utterances which vary in length depending on the protocol. The recordings were made using one of three types of microphones: a commonly available head-mounted noise-canceling microphone (Labtec LVA-8450), an array microphone (CNnetcom-Voice Array Microphone VA-2000), and a commonly available desktop far field microphone. The final corpus is sampled at 16 kHz at 16 bits per sample. Each audio file is accompanied by a word-level transcription. Corresponding information such as subject ID, age, sex, grade-level, and native language of speaker is also provided.

#### 3.2. University of Colorado Read and Summarized Stories Corpus

The CU Read and Summarized Stories Corpus (Cole and Pellom, 2006b) consists of transcribed speech data from 106 children in grades 3-5 within the Boulder Valley School District (Grade 3: 17 speakers, Grade 4: 28 speakers, Grade 5: 61 speakers) who read and summarized stories during a 30 min session from a set of 10 stories, and 221 children in grades 1 and 2 who summarized stories read to them from

Table 1  
Subject count in the CU Prompted and Read Children's speech corpus by grade level

Grade	K	1	2	3	4	5
#Children	84	136	150	92	91	110

a set of 62 stories. Third through fifth grade children also read 25 phonetically balanced sentences for future use in exploring strategies for speaker adaptation. Data were collected in a quiet room using a Labtec Axis-502 microphone. The data were recorded at 44 kHz and later re-sampled to 16 kHz for the purposes of experimentation. The current corpus consists of 10 different stories. The number of speakers per story is shown in Table 2. Each story contained 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. The additional children are used for acoustic model training in our current system together with the CU Read and Prompted Children's Corpus and the OGI Kids' Speech Corpus (Shobaki et al., 2000).

#### 4. Initial children's speech recognizer for oral reading recognition

Our initial work focused on improving performance of the SONIC speech recognizer for recognizing children's speech during oral reading (Pellom, 2001; Pellom and Hacıoglu, 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 Hidden Markov Models (HMMs). For the purposes of experimentation, children's acoustic models were estimated from 50 h of audio from the CU Read and Prompted Children's Speech Corpus, the OGI Kids' speech corpus grade K through 5 (Shobaki et al., 2000), and the first and second graders from the CU Read and Summarized Stories Corpus.

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 (van Vuuren et al., 2006). The reading tracking module determines the

Table 2  
Overview of 10 stories used in the CU Read and Summarized Story Corpus for (A) number of children who recorded the story, (B) number of words in the story, (C) number of unique words, (D) average summary length in words

Story #	(A)	(B)	(C)	(D)
1	22	572	205	150
2	22	532	207	129
3	12	932	364	181
4	12	1668	606	224
5	11	828	329	185
6	9	1078	389	161
7	8	1926	631	276
8	5	1157	526	133
9	3	933	460	101
10	2	919	417	90

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).

In 2003 we developed a baseline speech recognition system for oral reading recognition (Hagen et al., 2004). This initial 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, the following three sentences from an Interactive Book

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

were normalized as:

<s> IT WAS THE FIRST DAY OF SUMMervaca-  
TION </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 (i.e., no speech nor other text data) is used to construct the language model. Note that the sentence markers (<s> and </s>) are used to represent positions of expected speaker pause. Currently this baseline system is shown in Table 3(A) to produce a 16.5% word error rate. We evaluated using the 106 children in the CU Read and Summarized Story Corpus (Table 2).

##### 4.1. Improved speech recognition for oral reading recognition

In Section 1 we presented results from our pause analysis that showed children often do not pause at expected points of punctuation and frequently pause at wrong positions. In

Table 3  
Recognition of children's read aloud data

Experimental configuration	WER
(A) Baseline: single $n$ -gram language model	16.5%
(B) (A) + Begin/End sentence context modeling	13.0%
(C) (B) + between utterance word history modeling	10.8%
(D) (C) + dynamic $n$ -gram language model	9.8%
(E) (D) + VTLN	9.1%
(F) (E) + VTLN/SAT + SMAPLR (iteration 1)	7.9%
(G) (E) + VTLN/SAT + SMAPLR (iteration 2)	7.8%

order to further improve our speech recognition performance on read speech we extended our baseline system with new techniques that help to adjust to these phenomena.

#### 4.1.1. Improved sentence context modeling

Based on pause analysis, 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 ( $\langle s \rangle$  and  $\langle /s \rangle$ ). Results of this modification are shown in Table 3(B) and show a reduction in error from 16.5% to 13.0%.

#### 4.1.2. Improved word history modeling

Another observation is that 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 since the read aloud text segment is known to the recognition system, two consecutive words strongly indicate their successor in the text and therefore should be utilized by the system. That is, in our text example, the words “do today” should provide useful information to the recognizer that “Billy asked” may follow. Given that children who are actively learning to read may pause at any point in a text segment, we decided to modify the speech 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 13.0% to 10.8% as shown in Table 3(C).

#### 4.1.3. Dynamic $n$ -gram language modeling

During story reading we can also 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 considered estimating 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 weighted count 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 3(C)). At run-time, the recognizer implements a word-history buffer containing the most recent 15 recognized words. After decoding each utterance, the likelihood of the text within the word history buffer is computed using each of the position-sensitive language models. The language model with the highest likelihood is selected for the first-pass decoding of the subsequent utterance. This modification was found to further decrease the word error rate from 10.8% to 9.8% (Table 3(D)).

#### 4.1.4. Vocal tract normalization and acoustic adaptation

While the presented techniques effectively improve the reading recognition performance by advanced language modeling schemes, there still remains an acoustic gap that can be reduced. Therefore we further extend on our baseline system by incorporating the Vocal Tract Length Normalization (VTLN) method (Welling et al., 1999). Based on results shown in Table 3(E), we see that VTLN provides only a marginal gain (0.7% 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 et al., 2002) to yield a final recognition error rate of 7.8% absolute (Table 3(G)). By combining all of these techniques, we achieved a 52% reduction in word error rate relative to the baseline system. This work was later integrated directly within the COLit, 2004. In previous experiments (Hagen et al., 2003) we could show that when SAT, VTLN, and SMAPLR is already applied, the additional use of the presented read aloud directed language modeling techniques reduce the WER by about 20% relative (2% absolute). Therefore the impact of acoustic adaptation is approximately at the same level as the language modeling benefit.

## 5. High-accuracy recognition of oral reading based on subword units

To date, speech recognition systems designed to track speech as students read out loud present recognition results at the word-level; e.g., CMU’s Project LISTEN reading coach and our use of SONIC in the reading tracker used in Interactive Books in the Foundations to Literacy program. For children learning to read, our analyses of reading miscues suggest that whole-word modeling may not provide sufficient information to provide students with accurate and timely feedback based on their speech productions while reading out loud. For example, if the student says “sih” three times while reading the word “syllable,” it would be desirable for reading program to detect that the student is having difficulty sounding out this word, and then interact with the student to help them recognize the word. In order to better model events such as repeated syllables, partial words and mispronunciations, we propose to model words as comprised of constituent parts and analyze these constituents using a more fine grained approach.

### 5.1. Subword unit token flow architecture

In our subword unit based recognition system units can represent the word-initial, word-medial, and word-final phonetic unit sequences. Thus a subword unit represents a phoneme sequence. The phoneme sequence should be comprised of one or more phonemes and for every such

sequence there should exist a word that embeds it. The actual set of subword units is flexible and depends on the applied selection algorithm. For example, the word “destination” might be constructed from its subword pieces “(des)(ti)(na)(tion)” with corresponding subword phonetic sequences “(d eh s)(t ax)(n ey)(sh ax n)”. This type of representation can begin to account for partial word events (e.g., “des- destination”) or allow for novel methods for modeling pronunciation variabilities which can occur at the subword level. We point out that the determination of such subword units can, in principle, be determined either by hand or in a data-driven fashion. The latter realization makes such a modeling framework easily adaptable to languages other than English. Considering this representation of words, we extend on the notion of a single lexical prefix tree by considering a new architecture in which  $N$  lexical trees (where  $N = 3$ ) are connected in series and the original tree containing words in parallel, as shown in Fig. 1. Therefore the architecture evolves into a hybrid system that is able to recognize both, subword units or word-level units.

In this architecture, each lexical tree is used to encode a set of phonetic sequences representing portions of words (i.e., subwords). Each subword unit phonetic sequence is assumed to represent the beginning (Tree 1), middle (Tree 2), or end-part of a word (Tree 3). Using the token passing paradigm (Young et al., 1989), tokens flow into the initial tree modeling the beginning part of a word. Tokens propagate out of this network into Tree 2, which is used to model phonetic sequences from word medial positions. Finally tokens can pass from Tree 2 into Tree 3 in order to represent word-final phonetic sequences. The architecture allows for looping within Trees 1 and 2 to account for repeated instances of partial words (e.g., “it was the first day of sum- sum- summer vacation”) as well as looping within Tree 2 to account for words, which are modeled with a variable number of word-medial units. The search is made efficient due to the inherent ability to apply pruning

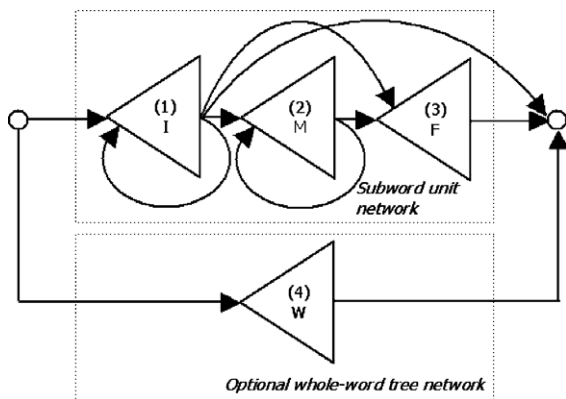


Fig. 1. Token flow-graph. A word in the subword unit network is modeled by one initial, an arbitrary number of medial, and one final subword unit tree. The loops in tree (1) and tree (2) account for repeated starts and multiple medial units per word. Thus, words comprised of any number of units can be represented.

and transitional constraints at the entry and exit points of each connected lexical tree as well as by applying transition costs estimated as subword unit  $n$ -gram probabilities. Tokens exiting each tree during the Viterbi search are inserted into a single word/subword event lattice. This word/subword event lattice can be processed to decode both a word sequence and, in some cases, sequences of partial word events. The final unit sequence is determined from the maximum likelihood path.

For the purposes of this work, we have extended the University of Colorado SONIC speech recognition system to incorporate the new lexical tree architecture. We point out that our final implementation correctly accounts for cross-unit triphone modeling within the search space. It should be noted that lexical prefix trees represent a compact representation of the phonemic sequences of words.

The hybrid system’s output is a sequence of subword units and words. Table 4 shows such a sequence that occurred while recognizing one of the speakers with the hybrid system. It can be seen that the speaker tried to pronounce the word WHISPERING. He pronounced the word only partially first, while the second attempt was a complete pronunciation, as the transcription “He imagined the people **whisper-** whispering to themselves” shows.

In (Bazzi, 2002) subword units are used for modeling out-of-vocabulary words for speech recognition. The authors use MIT’s Summit speech recognition system, which is based on finite-state transducers. Since the SONIC speech recognizer’s underlying architecture uses a lexical tree, as do most state-of-the-art systems, the techniques used in Bazzi’s system necessarily differ from this work. (Bazzi, 2002) also describes a subword unit selection algorithm, which is described later and compared to our approach.

## 5.2. Subword unit selection

In this article we propose a new algorithm for automatically determining a set of subword units given an input pronunciation lexicon consisting of words and phoneme sequences. First we describe two existing approaches proposed in (Creutz and Lagus, 2002 and Bazzi, 2002).

Table 4

Example sequence of words and subword units hypothesized by the recognizer

Word or unit	Start time (s)	Duration (s)
HE	225.05	0.75
IMAGINED	226.25	0.19
THE	226.44	0.51
PEOPLE	226.95	0.29
<b>WHIS-</b>	227.24	0.24
<b>PAXR-</b>	228.09	0.66
WHISPERING	229.05	0.15
TO	229.20	0.64
THEMSELVES	230.39	0.20

Start and time span of the words or units are shown.

Next we provide details of the proposed LZW-based method.

### 5.2.1. MDL-based subword unit selection

In (Creutz and Lagus, 2002) the authors introduce a subword unit selection algorithm for Finnish, using the Minimum Description Length (MDL) principle, which intends to find units, named morphs, in a purely data driven manner by optimizing a cost function representing the minimum description length of the encoded text as well as the resulting lexicon. This work was later extended to the Turkish language in (Hacioglu et al., 2003). The authors' goal is to minimize the entropy of the encoded text consisting of morphs and at the same time the number of bits required to store the codebook holding all discovered morphs. Therefore the cost function to be minimized is

$$C = \sum_{\text{tokens}} -\log(p(m_i)) + \sum_{\text{types}} k * l(m_i),$$

where a token represents a morph  $m_i$  in the encoded text sequence, all morphs in the codebook are called types,  $k$  specifies the number of bits needed to code a character,  $p(m_i)$  is the probability of  $m_i$  occurring in the text, and  $l(m_i)$  is the length of  $m_i$  in characters or phonemes, as it will be used in our case. The search for the optimal word segmentations is done recursively. The whole word is added as a morph to the codebook first and the cost function is evaluated, also every possible split of the word into to segments is evaluated, the best option is chosen and this process is repeated on the two segments until no further split is beneficial. Each time a word is analyzed it is first removed from the data structures such that a newly arisen split configuration can possibly be found. After every certain number of processed words the algorithm temporarily stops reading new words and loops over all already processed words trying to find a better segmentation which might have not been found earlier due to the earlier limited codebook size that is steadily increasing.

### 5.2.2. MI-Based subword unit selection

(Bazzi, 2002) describes a bottom-up clustering algorithm starting on phonemes as the basic unit. The algorithm concatenates units that co-occur most frequently. Mutual Information (MI) is used to determine which units should be merged in each iteration.

The initial unit set is represented by the phoneme set. For each pair of units occurring in the lexicon the units weighted mutual information is computed by

$$MI_w(u_1, u_2) = p(u_1, u_2) \log \frac{p(u_1, u_2)}{p(u_1)p(u_2)}$$

where  $p(u_1)$  and  $p(u_2)$  are the marginal probabilities of units  $u_1$  and  $u_2$  in the lexicon and  $p(u_1, u_2)$  is the relative frequency of the co-occurrence of  $u_1$  and  $u_2$ . In each iteration the mutual information is computed for each occurring pair of units and the top  $M$  pairs with the highest MI are merged in the lexicon and the unit set is extended with

these. In the rare case that one unit only occurs in the context of another unit the merged pair is added and the unit (or even both units) that only occurs in this context is removed from the unit set. Therefore the growth of the unit set after  $N$  iterations is close to but not necessarily  $N \times M$ .

### 5.2.3. Proposed LZW-Based subword unit selection

Our subword unit selection algorithm also aims at splitting words into sequences of phonetic units that occur most frequently in the lexicon under investigation. The algorithm performs the splitting in a data-driven and unsupervised manner. The derivation of the units is accomplished using a variation of the LZW (named after its inventors Lempel, Ziv, and Welch) text compression algorithm. This algorithm essentially determines frequent symbol sequences that can code a text sequence in a space-efficient manner. During unit selection only the encoding step is needed. A lexicon containing words and their respective pronunciations is fed to the LZW encoder. The algorithm creates a table of frequent units found from the phoneme sequences of each word in the lexicon. In order to facilitate the association of recognized subword units to whole-word sequences, we additionally mark candidate phoneme sequences (i.e., our subword units) as occurring at the beginning, middle, or end of a word.

The LZW encoding algorithm is changed slightly in our implementation. In the standard algorithm, after the input lexicon has been processed, the table contains units with various phoneme sequence lengths and word positions. In our application there exist different tables for different phoneme sequence lengths as well as for initial, medial, and end of word units. In addition to this multi-table implementation, we also record how often each phoneme string is looked up in the encoding tables. With this book keeping functionality it is possible to retrieve counts of how frequently a certain phoneme pattern occurs in a lexicon and therefore gives an indication of how likely it is to correspond to a reasonable subword unit. We restrict the bookkeeping to a maximum unit length of 4 phonemes.

Having these tables and the frequency counts it is possible to search through a phoneme sequence representing the pronunciation for a word and identify all possible split points. The tables, which are dependent on the number of phonemes and the units' position within the original word, are sorted by the count information. The higher the count, the higher in the list the corresponding unit will occur. The relative position of each subword unit in the list is an indication of how likely the unit represents a reasonable subword unit. We stress that the further use of the relative position for each unit in each table and the presence of multiple tables, dependent on unit length and within-word position, ensure the independence of the units' absolute frequencies of occurrence. Taking the absolute number would result in a preference for smaller units, ultimately phonemes, since these patterns occur most often. Therefore tables holding units with only a certain length and taking the relative rank of a unit in its table as its score enable

the discovery of even longer often-occurring phoneme patterns.

To split a word into subword units all possible subword sequences must be considered. For each of these paths a score is computed by averaging over the relative positions of the path's units in their corresponding lists. The path with the best score is chosen to represent the final unit sequence. To illustrate, let a word have the phoneme sequence  $p_1, \dots, p_n$ . There might be several possible unit sequences that result in the desired phoneme sequence when concatenated. The candidate unit sequences are denoted as  $u_{i1}, u_{i2}, \dots, u_{imi}$ . The score for each unit sequence  $i$  is given by

$$\text{Score}(i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \text{RP}(u_{ij}),$$

where  $\text{RP}(u_{ij})$  denotes the relative position of unit  $u_{ij}$  in its LZW-table. The unit sequence with the best average score represents the highest average position among all considered units and is therefore selected.

#### 5.2.4. Example subword unit representations

For experimentation we estimated subword units using a lexicon extracted from a typical 64k-word vocabulary for the US English Broadcast News Domain. For our lexicon containing 74k entries (64k + alternatives), the LZW method discovered 17,793 position dependent units, and the MDL algorithm 24,670. The MI algorithm can be stopped at any point and will produce a set of units representing the lexicon. The earlier it is stopped the smaller the set will be. We tested two result sets, after 200 iterations and after 1400 iterations. The unit set after 200 iterations (named MI200) had 5061 position dependent units and the set after 1400 iterations (named MI1400) resulted in 18,131 units. In comparison, the freely available NIST tool for US English syllabification results in 17,305 syllable-sized units. Leaving the position within the word (initial, medial, final) out of consideration the unit counts change to 14,529, 15,569, and 12,286 for LZW, MDL, and syllables, respectively. MI200 had 2051 and MI1400 had 13,499 units. In the following analysis we will focus more on MI1400, because the unit count is closer to the other algorithms' counts. Example split points for words are shown in Table 5 for the MDL, MI, and LZW algorithm.

#### 5.2.5. Evaluation of the different subword unit selection approaches

Clearly Table 5 suggests that there are some similarities and differences between the three algorithms. Next, we compare the statistical properties of each set of subword units with respect to syllables. Syllables seem to be a reasonably fine grained basic unit for recognition. Generally words can be represented by syllables and many non-words that might be due to mispronunciations and incomplete utterances. For English, rule-based software tools are available to segment words into syllables, this might not be true for any given language and therefore purely data

Table 5  
Word and corresponding subword units

ACTING	LZW	(ae kd t) (ix ng)
	MDL	(ae kd t ix ng)
	MI1400	(ae kd t ix ng)
ANYONE	LZW	(eh n iy) (w ax n)
	MDL	(eh n iy) (w ax n)
	MI1400	(eh n iy) (w ax n)
ASTONISHING	LZW	(ax s t) (aa n) (ix sh) (ix ng)
	MDL	(ax s t aa n ix sh) (ix ng)
	MI1400	(ax s t aa n ix sh ix ng)
BELIEVE	LZW	(b ax) (l iy v)
	MDL	(b ax l iy v)
	MI1400	(b ax l iy v)
CARRYING	LZW	(k eh r iy) (ix ng)
	MDL	(k eh r) (iy ix ng)
	MI1400	(k eh r iy) (ix ng)

driven approaches are of great benefit. Finally we evaluate the methods within the context of an interactive literacy tutor for children.

**5.2.5.1. Unit length distribution.** Table 6 provides a comparison of the word segmentation algorithms' units and the NIST syllabification software output in terms of the mean and standard deviation of subword unit length (computed based on phoneme count). We can see that the proposed LZW method produces units which are closer in size distribution to syllables than the other algorithms' outputs.

**5.2.5.2. Histogram analysis.** We have computed the unit length histogram for each of the methods described earlier. Fig. 2 shows the histograms. It can be seen that LZW and syllables again show very similar properties.

We use the Chi-square ("Goodness of Fit") test to measure the distance between histograms shown in Fig. 2. The distance is given by

$$\chi^2 = \sum_{i \in \text{LEN}} (p(i) - h(i))^2 / h(i),$$

where  $p$  and  $h$  each denote a histogram and LEN denotes the set of relevant unit lengths. Based on this analysis we found the goodness of fit between MDL and syllables to be 0.44, between MI1400 and syllables to be 0.83, compared to 0.25 for the proposed LZW and syllable units. MI200 fits syllables by a goodness of fit value of 0.28, which is close to LZW's match, but the number of units for MI200 (2051) is much smaller than the number of syllables or LZW's unit set (both more than 12,000). Therefore the LZW method provides units which share more similar statistical features with syllables than the other algorithms.

Table 6  
Mean number of phonemes and standard deviation ( $\mu/\sigma$ ) for the different sets of units

MDL	MI200	MI1400	LZW	Syll.
4.52/1.82	3.38/1.32	5.07/1.98	3.29/0.72	3.52/0.87

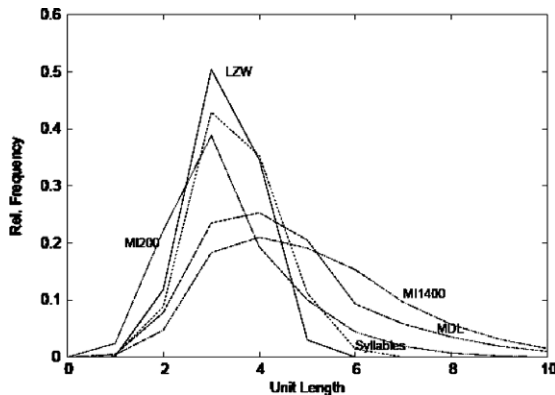


Fig. 2. Histograms for the unit sets found by MDL, MI, LZW, and the NIST syllabification tool.

**5.2.5.3. Unit to syllable correlation.** Another indication how close LZW, MDL, or MI based units are to syllables is the fraction of discovered units per word that actually appear as a syllable when using the NIST syllabification tool. The analysis is performed as follows. A word is segmented into units by LZW, MI, and MDL, and into syllables by the NIST tool. The NIST tool outputs at least one syllabification, sometimes even three or more alternatives. For a word it is checked for each unit found by one of the data driven methods if it appears as a syllable for that word in the NIST tool's output. The evaluation using a typical Broadcast News lexicon of 64k words shows that in average 74% of the units discovered by the LZW based method occur as syllables compared to only 41% for the MDL based method and 54% by MI1400. MI200 overlaps with syllables by 68%, which is relatively high, but still the number of units indicates significant differences to syllables. Furthermore MI generates many single-phoneme units. Therefore in addition to the statistical similarity seen in earlier sections, also the phonemic similarity between the proposed LZW units and syllables is more significant compared to the other algorithms' results.

### 5.3. Subword unit language modeling

Language modeling is a critical component in the subword unit recognition framework. We first transform source text into sequences of subword units. Since whole-words often have pronunciation alternatives, we must take this into consideration during subword unit  $n$ -gram estimation. To account for the correct handling of whole-word pronunciation alternatives, a new language modeling tool was developed to compute language models in terms of subword units. Details can be found in (Hagen and Pellom, 2005a,b).

### 5.4. Subword unit recognition evaluation

#### 5.4.1. Experimental data

We evaluate the new architecture in the framework of an interactive literacy tutor where children are reading text

out loud. For all experiments we use the 106 speaker corpus described in Section 3.2 as a test set. Each of the 106 speakers reads one out of ten stories. The average length of a story is 1054 words. As in Section 4, the training data consists of 50 h of audio from the CU Read and Prompted Children's Speech Corpus, the OGI Kids' speech corpus grade K through 5, and the first and second graders from the CU Read and Summarized Stories Corpus.

#### 5.4.2. Hybrid Word/Subword unit based system

The proposed recognition system outputs subword units as well as whole-words. The subword units that can be recognized are computed prior to the recognition process by the subword unit selection algorithm presented earlier. For the test corpus, the total vocabulary of all stories consists of 1789 words. The LZW-based unit selection algorithm transforms these words into 1052 word-initial units, 503 word-medial and 646 word-end units.

We compare the word based and the hybrid system in terms of word error rate and real-time recognition factor. For assessment, the word/subword unit sequences are transformed into pure word sequences by simply traversing through the word/unit sequence and looking up nearest matching words for the unit sub-sequences. Table 7 gives an overview of the simulations.

The results in Table 7 suggest that the proposed architecture provides *equally* accurate recognition of the test corpus as the baseline whole-word system, which uses a single lexical tree. The system based on MDL, MI, and syllables had comparable performances to the LZW system, although a marginal WER advantage of the syllable and LZW based systems can be noticed. We also considered constructing the hybrid word/subword network into a single lexical tree. However, the first pass word error rate (WER) here increased slightly to 10.0%. This difference was minimal compared to the increase in the real-time factor which reached 0.64 and was therefore more than 13% relative worse than the proposed system's real-time behaviour. The proposed method additionally enables the distinct pruning of unit tokens, leaving word tokens to the previous wider pruning beam. Therefore the hybrid system is as accurate and almost as fast as the whole-word based system, but additionally allows partial words to be detected.

Table 7

Word error rate (WER) and real-time factors (RTF) for various system configurations

Configuration	WER	RTF
Word based, 1 lexical tree	9.9%	0.56
Word/LZW-subword, 1 lexical tree	10.0%	0.64
MDL hybrid system	10.1%	0.60
MI1400 hybrid system	10.2%	0.59
Syllable hybrid system	9.9%	0.59
<b>Proposed LZW hybrid system</b>	<b>9.9%</b>	<b>0.57</b>

Results are shown without acoustic adaptation.

Our investigations show that the special lexical tree layout slightly decreases (or does not affect) the WER and more drastically changes the real-time factor in a positive way. These differences in real-time can be interpreted in the following way. The word-based system is expected to be the fastest one because the total number of words in its lexical tree is low compared to the hybrid systems' number of words and subwords, and therefore the search should be most efficient. The architectural change which is able to overcome this behaviour to a large extent, is the distribution of words and subword units to different lexical trees and the control of the transitions from one tree to another, as shown in Fig. 1. This token flow control restricts the number of possibilities of words and subword units following each other and therefore speeds up the recognition process.

The real benefit of this new hybrid system appears to be the partial word detection capability. The partial word detection system looks for cases where a word consisting of subword units was not completed in the recognizer's hypothesis. The hypothesis containing subword units is transformed into a word sequence, where potentially partial words are marked. By comparing the hypothesis with the transcript the correct detections and false alarms can be identified. Fig. 3 shows the hybrid systems' ROC curves for partial word detections.

For partial words with a significant fraction of the word present in the speech signal, namely three spoken letters or more, the syllable based system gives the best results, followed by the LZW system. The notion of letters in this context is not precise, still the speech data available for this study transcribed partial words using letters. This induces some vagueness, but still provides a notion for the approximate length of the sounded out part of the word. The use of MDL and MI1400 units results in a significantly worse partial word detection versus false alarm rate performance. The ROC analysis is limited to false alarm rates less than one percent, since too many wrong mispronunciation alarms will lead to a frustrating experience for the user. While it is still debateable if the syllable is the best subword

unit representation, this result seems to support this claim. LZW based units have similar statistical features to syllables, as shown in Section 5.2.5, and not surprisingly show similar partial word detection results. It is interesting to see that MDL and MI1400 based units share less statistical similarities with syllables and at the same time lack partial word detection performance.

## 6. Discussion

In this article we have focused on high-accuracy children's speech recognition in the context of an interactive literacy tutor. In particular this work considered modeling strategies for oral reading recognition. We presented special language modeling techniques that were motivated by our experience from listening to children during this task. Children tended to ignore sentence boundaries or paused at positions in the text where no punctuation indicated any stops. Dynamic language modeling, sentence context modeling, and word history modeling helped to improve the recognizer's ability to keep track with children reading over sentence boundaries and recover from pauses the recognizer could not anticipate based on the story text. The improvement in accuracy from language modeling techniques alone was very high at 38% relative. We also found that at a higher grade, children tend to read over punctuation marks more often. This might be due to improved word reading skills in older children that read faster but do not have the ability to comprehend the text well enough to produce an appropriate prosodic structure while they read aloud.

This discussion shows that accurate language modeling provides a critical contribution to accurate recognition performance in the read aloud task. Prior work on children's speech recognition has shown that recognition of children's speech is difficult due to acoustic as well as language variations. Fortunately, the read aloud task forces the language uncertainty to be reduced due to the reading of known texts. This enables accurate modeling of the language, taking out the uncertainty about the language to a high degree, and therefore enabling accurate recognition compared to other children ASR tasks, where knowledge of what the child is supposed to speak is not available.

Our work since 2003 has consistently shown that finer grained analysis is important for interactive reading tutors. A standard recognizer might ignore partial words and therefore not enable detection of possible reading problems, or the recognizer might substitute another word for the partial word indicating stronger difficulties than are actually present. A subword unit recognition system also has the potential of detecting pronunciation difficulties. In reading instruction, the automated reading tutor must be able to detect significant deviations from expected pronunciations in order to provide the student with accurate feedback about their reading performance. Whole word approaches produce a single score for each word, and

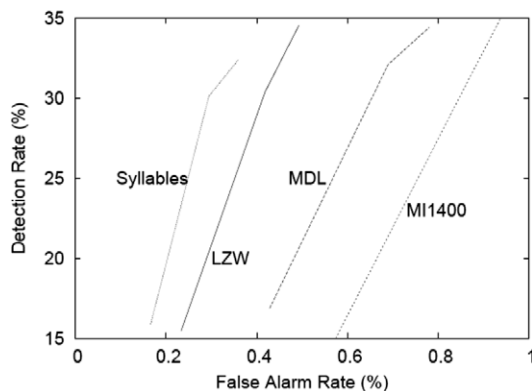


Fig. 3. ROC curves for the hybrid systems based on syllables, LZW, MDL, and MI1400 units.

cannot provide the student with important feedback about which syllable in a word has been mispronounced.

One main focus in the design of the subword unit system was competitive word accuracy and speed with the word-based system. This allows the system's justification for use in real environments where word accuracy is the first main focus, and special event detection is a valuable feature. The solution required a hybrid system that enables recognition of both words and subword units automatically, with the appropriate unit selected solely on the basis of the recognizer's internal search scores during the recognition process. It was interesting to observe that the system behaves in a way that is consistent with human perception while listening to a student read out loud—it recognizes whole-words while the student reads well, and appears to switch to subwords only when the student is producing partial words and similar effects. This is a desirable emergent property of the system, which was not designed to behave in this manner.

## 7. Conclusion

In this work we presented various approaches to facilitate reading tracking in automated reading tutors. Some early analysis showed that multiple acoustic models, each trained on data from a specific grade level, are not able to outperform one VTLN normalized model using all available training data. Dynamic language modeling, cross utterance context modeling, and word history modeling have previously provided significant speech recognition accuracy and performance improvements within the read aloud task. To be able to analyze special events, i.e., mispronunciations and partial words, a hybrid word/subword unit system was proposed. The system has competitive accuracy and performance to a word based system but is able to recognize speech on the subword level. The switch between word and subword unit recognition is done on the fly and has the distinct advantage of allowing for the detection of partial words.

## Acknowledgements

This work was supported by grants from the National Science Foundation's ITR and IERI Programs under grants NSF/ITR: IIS-0325399, NSF/ITR: REC-0115419, NSF/IERI: EIA-0121201, NSF/ITR: IIS-0086107, NSF/IERI: 1R01HD-44276.01; and Dept. of Education grant IES-R305G040097. The authors would like to thank Linda Corson, David Cole, Scott Schwartz and Taylor Struempfler for collection and transcription of the CU children's speech corpora, and John Paul Hosom at the Center for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute (OGI) for his collaboration in developing the data collection protocols for the CU Prompted and Read Speech Corpus. The authors would also like to acknowledge several researchers at the Helsinki University of Tech-

nology, especially Mikko Kurimo, Vesa Siivola, Krista Lagus, and Mathias Creutz for their earlier work and inspiration related to data-driven subword unit modeling for speech recognition within the Finnish language.

## References

- Aist, G., Chan, P., Huang, X., Jiang, L., Kennedy, R., Latimer, D., Mostow, J., Yeung, C., 1998. How effective is unsupervised data collection for children's speech recognition? In: Proc. ICSLP 98 Sydney, Australia.
- Arcy, S., Wong, L., Russel, M., 2004. Recognition of read and spontaneous children's speech using two new corpora. In: Proc. ICSLP 2004, Jeju Island, Korea.
- Banerjee, S., Beck, J., Mostow, J., 2003a. Evaluating the effect of predicting oral reading miscues. In: Proc. Eurospeech 2003, Geneva, Switzerland.
- Banerjee, S., Mostow, J., Beck, J., Tam, W., 2003b. Improving language models by learning from speech recognition errors in a reading tutor that listens. In: Proc. Second Internat. Conf. on Applied Artificial Intelligence 2003, Fort Panhala, Kolhapur, India.
- Bazzi, I., 2002. Modelling out-of-vocabulary words for robust speech recognition. Ph.D. Thesis, MIT, June 2002, Department of Electrical Engineering and Computer Science.
- Cole, R., Hosom, P., Pellom, B., 2006a. University of Colorado Prompted and Read Children's Speech Corpus. Technical Report TR-CSLR-2006-02, Center for Spoken Language Research, University of Colorado, Boulder.
- Cole, R., Pellom, B., 2006b. University of Colorado Read and Summarized Stories Corpus. Technical Report TR-CSLR-2006-03, Center for Spoken Language Research, University of Colorado, Boulder.
- Cole, R.A., Van Vuuren, S., Pellom, B., Hacıoglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., Yan, J., 2003. Perceptive animated interfaces: first steps toward a new paradigm for human-computer interaction. Proc. IEEE: Special Issue on Human-Computer Multimodal Interface 91 (9), 1391–1405.
- Cole, R., Wise, B., Van Vuuren, S., 2006. How Marni teaches children to read. *Educ. Technol.* 47 (1), 14–18.
- COLit, 2004. Colorado Literacy Tutor Project. <<http://www.colit.org>>.
- Cosi, P., Pellom, B., 2005. Italian Children's speech recognition for advanced interactive literacy tutors. In: Proc. Eurospeech 2005, Lisbon, Portugal.
- Creutz, M., Lagus, K., 2002. Unsupervised discovery of morphemes. In: Proc. Workshop on Morphological and Phonological Learning of ACL-02, Philadelphia, pp. 21–30.
- Das, S., Nix D., Picheny, M., 1998. Improvements in children's speech recognition performance. In: Proc. ICASSP 98, Seattle, WA.
- Eskenazi, M., 1996. KIDS: A database of children's speech. *J. Acoust. Soc. Amer.* 100 (4, Part 2).
- Fogarty, J., Dabbish, L., Steck, D.M., Mostow, J., 2001. Mining a database of reading mistakes: For what should an automated Reading Tutor listen? In: Proc. Tenth Internat. Conf. on Artificial Intelligence in Education (AI-ED) 2001, San Antonio, Texas.
- Gales, M., 1997. Maximum likelihood linear transformations for HMM-based speech recognition. Technical Report, CUED/F-INFENG/TR291, Cambridge University.
- Giuliani, D., Gerosa, M., 2003. Investigating recognition of children's speech. In: Proc. ICASSP 2003, Hong Kong.
- Gustafson, J., Sjolander, K., 2002. Voice transformations for improving children's speech recognition in a publicly available dialogue system. In: Proc. ICSLP 2002, Denver, Colorado.
- Hacıoglu, K., Pellom, B., Ciloglu, T., Ozturk, O., Kurimo, M., Creutz, M., 2003. On lexicon creation for Turkish LVCSR. In: Proc. Eurospeech 2003, Geneva, Switzerland.
- Hagen, A., Pellom, B., 2005a. A Multi-layered lexical-tree based token passing architecture for efficient recognition of subword speech units. In: The 2nd Language and Tech. Conf., Poznan, Poland.

- Hagen, A., Pellom, B., 2005b. Data driven subword unit modeling for speech recognition and its application to interactive reading tutors. In: *Interspeech 2005*, Lisbon, Portugal.
- Hagen, A., Pellom, B., Cole, R., 2003. Children's speech recognition with application to interactive books and tutors. In: *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, St. Thomas.
- Hagen, A., Pellom, B., Van Vuuren, S., Cole, R., 2004. Advances in children's speech recognition within an interactive literacy tutor. *HLT-NAACL*, Boston, May 2004.
- Lee, S., Potamianos, A., Narayanan, S., 1997. Analysis of children's speech: duration, pitch and formants, In: *Proc. EUROSPEECH 97*, Rhodes, Greece.
- Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Amer.* 105, 1455–1468.
- Lee, K., Hagen, A., Romanyshyn, N., Martin, S., Pellom, B., 2004. Analysis and detection of reading miscues for interactive literacy tutors. *COLING*, Geneva, Switzerland.
- Li, Q., Russell, M., 2002. An analysis of the causes of increased error rates in children's speech recognition. In: *Proc. ICSLP 02*, Denver, Colorado.
- McCandless, M., 1992. Word rejection for a literacy tutor. S.B. Thesis, MIT, May 1992, Department of Electrical Engineering and Computer Science.
- Mostow, J., Roth, S.F., Hauptmann, A.G., Kane, M., 1994. A prototype reading coach that listens. In: *Proc. of AAAI-94*, Seattle, WA, pp. 785–792.
- Mostow, J., Beck, J., Winter, S., Wang, S., Tobin, B., 2002. Predicting oral reading miscues. In: *ICSLP 2002*, Denver, Colorado.
- Pellom, B., 2001. SONIC: The University of Colorado Continuous Speech Recognizer. Technical Report TR-CSLR-2001-01, University of Colorado.
- Pellom, B., Hacioglu, K., 2003. Recent improvements in the CU SONIC ASR system for noisy speech: the SPINE task. In: *Proc. ICASSP 2003*, Hong Kong.
- Potamianos, A., Narayanan, S., 2003. Robust recognition of children's speech. *IEEE Trans. Speech Audio Process.* 11, 603–616.
- Potamianos, A., Narayanan, S., Lee, S., 1997. Automatic speech recognition for children. In: *Proc. EUROSPEECH 97*, Rhodes, Greece.
- Shobaki, K., Hosom, J.P., Cole, R., 2000. The OGI Kids' Speech Corpus and recognizers. In: *Proc. ICSLP 2000*, Beijing, China.
- Siohan, O., Myrvoll, T., Lee, C.H., 2002. Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer, Speech and Language* 16, 5–24.
- Spache, G.D., 1981. *Diagnostic Reading Scales*. Del. Monte Research Park, Monterey, CA 93940: CTB, Macmillan/McGraw-Hill.
- Tam, Y.C., Mostow, J., Beck, J., Banerjee, S., 2003. Training a confidence measure for a reading tutor that listens. In: *Proc. Eurospeech 2003*, Geneva, Switzerland.
- van Vuuren, S., Cole, R., Ngampatipatpong, N., 2006. Providing feedback to students while reading out loud in interactive books. Technical Report TR-CSLR-2006-01, Center for Spoken Language Research, University of Colorado, Boulder.
- Welling, L., Kanthak, S., Ney, H., 1999. Improved methods for vocal tract length normalization. In: *Proc. ICASSP 99*, Phoenix, Arizona.
- Wise, B., Cole, R., Van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., Tuantranont, J., Pellom, B., 2005. Learning to read with a virtual tutor: foundations to literacy. In: Kinzer, C., Verhoeven, L. (Eds.), *Interactive Literacy Education: Facilitating Literacy Environments through Technology*. Lawrence Erlbaum, Mahwah, NJ.
- Young, S.J., Russell, N.H., Thornton, J.H.S., 1989. Token passing: a simple conceptual model for connected speech recognition systems. Cambridge University, Technical Report CUED/F-INFENG/TR.38.