Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 861–873
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

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%.

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 (A. Hagen), pellom@cslr.colorado.edu (B. Pellom), cole@cslr.colorado.edu (R. Cole).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.05.004

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 Hacioglu, 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-