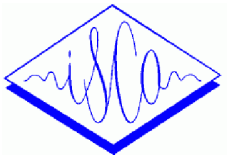

Second Workshop on Child
Computer Interaction (WOCCI 2009)
Cambridge, MA, USA
November 5, 2009

A Reference Verification Framework and its Application to a Children's Speech Reading Tracker

[Extended Abstract]

Daniel Bolaños
Boulder Language Technologies
2960 Center Green Court, Suite 200
Boulder, Colorado, USA
dani@bltek.com

Wayne H. Ward
Boulder Language Technologies
2960 Center Green Court, Suite 200
Boulder, Colorado, USA
wward@bltek.com

Ronald A. Cole
Boulder Language Technologies
2960 Center Green Court, Suite 200
Boulder, Colorado, USA
rcole@bltek.com

ABSTRACT

In this article we present a novel approach to reference verification, the problem of determining if a speakers' utterance matches a specificied reference (text) string, and discuss its application to a reading tracker system for children's speech.

Unlike other reading tracker systems proposed in the literature that are built over conventional speech recognizers with ad-hoc language models, the reading tracker described here is designed specifically for the task of estimating whether a child has read an expected sequence of words out loud; the tracker is designed to deal in a natural and flexible way with disfluencies that frequently appear in children's speech while reading out loud, (e.g., partial-words, repetitions, self-corrections, etc), and to overcome problems caused by using language models within the reference verification task. Two mechanisms have been introduced for this purpose, the utilization of filler models and the inclusion of backward inter-word transitions in the decoding network.

While this article focuses on the approach used to overcome errors observed in previous systems, the performance of this system will be evaluated on a corpus of children's speech while reading out loud and compared to the performance of a "traditional" reading tracker system that are built on top of a speech recognition system. The results of this comparison will be presented at WOCCI 2009.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*

General Terms

Algorithms

ICMI-MLMI'09 Workshop on Child, Computer and Interaction November 5, 2009, Cambridge, MA, USA

Keywords

Reference verification, children's speech, reading tracker

1. INTRODUCTION

Reference verification is a general task consisting of determining whether an interval of speech corresponds to a reference sequence of lexical unit, such as a text string. Typical scenarios in which reference verification plays a key role are pronunciation assessment or reading tracking tools [1, 3].

In this article, a general purpose reference verification framework is described and utilized in the context of a standalone reading tracker system. This system does not make use of any language model but rather used a decoding network specifically designed to deal with reading disfluencies that commonly appear in children's speech in a flexible way. Additionally, the proposed system reduces the overhead and complexity of using a large vocabulary continuous speech recognizer with an ad-hoc (possibly dynamic) language model.

The system is composed of two modules: the first module, based on a static decoding network specifically designed to deal with disfluencies, produces a hypothesis. The second module computes confidence values for each of the lexical units in the hypothesis and makes the final decision as to whether the hypothesized words are correctly read.

2. REFERENCE VERIFICATION FRAMEWORK

In this section we describe the architecture of the reference verification system. The input to the system is a reference string containing a sequence of lexical units that the user is prompted to read and the corresponding uttered interval of speech. The output is a classification of those units as correctly or incorrectly read.

2.1 Static decoding network for reference verification

Initially, a static decoding network is built from the lexical units contained in the reference string and the set of 3-state triphone HMMs. In this network, unlike a conventional decoding network, only transitions between lexical units as they appear in the reference string are allowed.

Additionally, optional silence and filler models are added in between each pair of lexical units. The filler model, con-

sisting of one self-transitioning HMM-state, is intended to account for speech frames corresponding to reading disfluencies, such as hesitations, partial words or self-corrections. These types of disfluencies are commonly in children's speech while reading out loud and sounding out words and therefore must be modeled explicitly. We note that if the reference verification task is intended for isolated speech (e.g., in assessing word recognition skills) neither the silence nor the filler models are needed.

In the case of a reading tracker, using a filler model is necessary for dealing with partial words or word mispronunciations. However, repetitions or sentence restarts, also quite common in children's speech, can be explicitly modeled by allowing backward transitions in the network. In the case of words, these transitions are expressed by backward pointers from the final node[1] of each word to the initial node of any of the preceding words in the current sentence and the word itself. This way, whenever a sentence restart occurs, the system is able to modify its state accordingly, thus easing the work of the filler model.

Once the decoding network is built, it is optimized by using a forward-backward node merge process similar to the one presented in [5]. Since admissible pronunciations of a given lexical unit typically share a number of phones, this simple mechanism allows a considerable reduction of the network size. This allows reducing the memory requirements and, most importantly, speeds-up the search.

2.2 Rejection module

Given that the decoding network previously introduced acts similarly to a forced alignment mechanism (with the exception of the filler models and the backward pointers), it is desirable to make use of a rejection module that, using the word and phone-alignments present in the hypothesis, is able to classify the reference words as correctly or incorrectly read. In this case, word-level confidence estimates are obtained using a set of Support Vector Machine (SVM) classifiers.

3. PROVIDING FEEDBACK TO THE USER

In the proposed reading tracker, designed to provide feedback to users on the accuracy and fluency of their speech while or shortly after reading aloud, an application interface layer supports communication between the user and reference verification system. This application highlights the sentence in the text the reader should read aloud and highlights the text either while it is read by the user or presents feedback on the child's performance after a sentence or paragraph has been read. Feedback can be provided by the system in these two ways, as follows:

- Word-level feedback: under this working mode the system highlights the text word by word as the words are read by the user. This can be done by retrieving partial hypothesis every specified (short) period of time; if the best partial path ending at the current time frame contains the immediately following word in the text, the word is highlighted, otherwise no feedback is provided in the expectation that the next best partial path retrieved will contain the word.

[1]Here the term node is used instead of state since not all the nodes in the network represent HMM-states.

- Utterance-level feedback: in some cases, providing word-level feedback as the child reads may be a source of distraction since the child may become overly conscious of it. We have observed, for example, that if the reading tracker fails to highlight a word, the child will pause and wait for the reading tracker to catch up. For this reason, it may be preferable to provide feedback at the end of a sentence, paragraph or page, and provide the child means for reviewing his or her performance.

Additionally, the user interface allows the reader to listen to the correct pronunciation of any word, produced by a lifelike computer character with accurate visual speech in the text just by clicking on it. Other interesting features are also available, like allowing the child to repositioning the cursor at the beginning of the current sentence.

4. SYSTEM EVALUATION

The performance of the proposed system will be evaluated in comparison with an existing benchmark system similar to [3], that makes use of Sonic [4] and dynamic language models, and a syllable-lattice based system [1].

The performance will be evaluated using the Classification Error Rate (CER) defined as the percent of words in the reference text that have been correctly tagged, as present (correctly) or absent (incorrectly) read, by the system. The reference classifications are generated by aligning the reference (prompt) string against hand generated transcriptions of the corresponding speech. Each word in the reference string aligned with the same word in the hand transcription is marked as present (read correctly). Words in the reference not aligned with the same word in the hand transcript are marked as absent (not read correctly).

The speech corpora used for the evaluation is [2].

5. ACKNOWLEDGMENTS

This research was supported by a grant from National Institute of Child Health and Development (NICHD) grant 1 R44 HD055028-01 to Mentor Interactive Incorporated (R. Cole, PI). The views expressed in this article do not necessary represent the views of the NICHD.

6. REFERENCES

[1] D. Bolaños, W. Ward, S. V. Vuuren, and J. Garrido. Syllable lattices as a basis for a children's speech reading tracker. In *Proceedings of Interspeech 2007*, pages 198–201. ISCA, August 2007.

[2] R. Cole and B. Pellom. University of colorado read and summarized story corpus. Technical report, University of Colorado, March 2006.

[3] A. Hagen, B. Pellom, and R. Cole. Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication*, 49(12):861–873, December 2007.

[4] B. Pellom. The university of colorado continuous speech recognizer. Technical report, University of Colorado, March 2001.

[5] J. Shao, T. Li, Q. Zhang, Q. Zhao, and Y. Yan. A one-pass real-time decoder using memory-efficient state network. *IEICE - Transactions on Information and Systems*, E91-D(3):529–537, March 2008.