Combining Acoustic and Language Model Miscue Detection Methods for Adult Dyslexic Read Speech

Morten Højfeldt Rasmussen[1], *Børge Lindberg*[1], *Zheng-Hua Tan*[1]

[1]Department of Electronic Systems, Aalborg University, 9220 Aalborg Ø, Denmark
{mr,bli,zt}@es.aau.dk[1]

Abstract

One important feature of automatic reading tutors is their ability to detect miscues in order to provide feedback to the student and/or to automatically evaluate the student's reading proficiency. The focus of this paper is on improving accuracy of detecting miscues in dyslexic read speech. We present a miscue detection method that combines a specialized language model and the goodness of pronunciation (GOP) score. The language model is augmented with a subset of the real word substitutions that are observed in the training set. Experiments have been conducted on a corpus containing adult dyslexic read speech. At a miscue detection rate of 34% the false rejection rate (FRR) using only the specialized language model is 2.6%, for the GOP score it's 3.4%, whereas for a combination of the two miscue detection methods the FRR is only 1.8%, which is a 31% relative improvement of FRR when compared to the specialized language model method.

Index Terms: miscue detection, automatic reading tutor, dyslexia, read speech

1. Introduction

One important feature of automatic reading tutors is their ability to detect miscues in order to provide feedback to the student and/or to automatically evaluate the student's reading proficiency. Miscues are usually detected using an automatic speech recognizer (ASR) using specialized language models, [1], [2], sometimes in combination with confidence measures, [3].

In this paper we combine a specialized language model with the goodness of pronunciation score to detect reading miscues. We augment the language model with a subset of the real word substitutions that are observed in the training set which comes from a corpus of adult dyslexic read speech (see Section 4). The hypothesis is that the optional word substitutions in the language model will detect some of the substitution mistakes made by the reader, and that the GOP score will detect miscues that are phonetically different from the target words (the words in the target/prompt text).

The rest of the paper is organized as follows. Section 2 describes dyslexics' reading miscues. Section 3 presents the proposed miscue detection approach. Section 4 gives an overview of the corpus of dyslexic read speech. Section 5 describes the evaluation setup. Section 6 presents the results and discusses. Section 7 concludes.

2. Reading miscues

Miscues occur on either the word or sentence level. When dyslexics read a text out loud they produce a wide variety of miscues. Some of these miscues are listed here:

- Word level
  - real word substitution
  - Nonsense-word substitution
  - Inflection mistake
  - Omission and insertion of letters
  - Substitution of letters
  - Reversal of letter order
  - Elongation of consonants/vowels
  - Partially read words
  - Mispronunciations
  - Split words
  - Unnatural prosody
- Sentence level
  - Omission and insertion of words
  - Repetition of words
  - Jumping back more than one word
  - Long pause between words

Modeling these miscues well is necessary in order to produce a highly accurate miscue detection system. This, however, is a challenging task, as some of the listed miscues vary extensively in realizations, e.g. mispronunciations can be virtually any change of the target word at the phonetic level.

The work presented here is the first step in trying to improve the miscue detection accuracy of the system presented in [4]. We focus on combining two detection methods. One is to model real word substitutions, as [5] indicates that the miscue predicting power is high when modeling these miscues for a small closed vocabulary. The other is using the GOP score on the word level in order to detect miscues that are not modeled by the word substitutions.

3. Detecting miscues

To detect reading miscues is different from detecting the words in the target text. Since the target words and their order is known it is straight-forward to construct a language model (LM) that models a correctly read target text. As miscues are introduced, however, the simple LM will not suffice. There are a number of places in an automatic reading tutor where miscue detection can be implemented, for example in connection with the:

- language model,
- lexicon/dictionary,
- acoustic models,

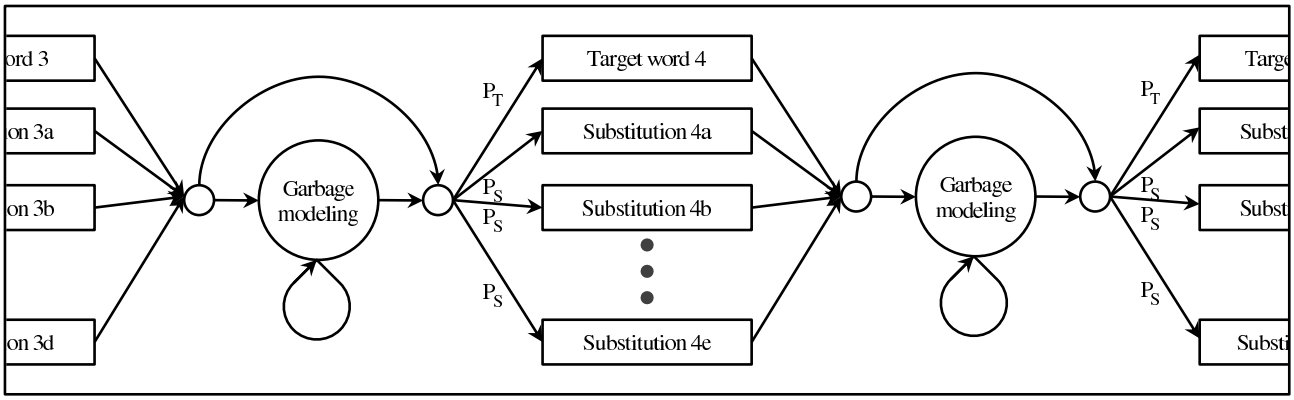

Figure 1: *A part of the specialized language model centered on the fourth target word, which in this case has 5 (a,b, ... , e) optional substitutions. P_T is the probability of the target word, P_S is the probability of a substitution.*

- tracking method.

The presented miscue detection methods are implemented in connection with the language model (explicit modeling of substitution miscues) and the acoustic models (GOP). An example of detecting miscues by means of the lexicon is by adding transcriptions of false starts and near misses [6]. An example of detecting miscues by means of tracking is by aligning the output of the recognizer to the target text [6]. If an aligned recognized word and target word are not the same, a miscue is detected.

3.1. Base setup

The base of the miscue detection system consists of a speech recognizer with a forced-alignment type language model. The language model is constructed from the target text with optional garbage modeling between words. The garbage modeling consists of a general phone model, a model of speaker noises (noises coming from the speaker's mouth), a model of intermittent noises (from the environment), a model of stationary noises (continuous noise from e.g. a fan), and a silence model. The language model cannot detect any miscues by itself as the recognizer will be forced to align each target word to part of the speech. One strength of this base setup is its very low false rejection rates, although at low miscue detection rates, as all words are being recognized by this setup. We then add miscue detection methods to this base and start detecting miscues (at the cost of false rejections).

3.2. Explicit substitution modeling

Miscue detection accuracy can be improved by explicitly modeling expected reading miscues. One way of doing this is by adding substitution miscues found in a training set as alternative words in the language model as done in [7]. They select likely substitution miscues from a large database of oral reading miscues developed by Richard Olson, Helen Datta, and Jacqueline Hulslander at the University of Colorado. [7] tried out different selection criteria, e.g. for each target text word selecting the top m most frequent miscues or the miscues that at least n students uttered. Because the text corpus we are using is small in comparison, where most miscues occur only once or twice, we needed a different selection criterion, and including all observed miscues in the language model results in very high false rejection rates (see Section 6). We therefore only select substitutions that do not look like short partial instances of the target word. In this way we reduce the number of false rejections by e.g. recognizing the target word "*icicle*" as a short substitution word like "*I*" followed by garbage models. We exclude a substitution when the following two conditions are met:

$$\frac{[longest\ common\ phone\ sub-sequence]}{[\#\ phones\ in\ substitution]} > \alpha \quad (1)$$

$$\frac{[\#\ phones\ in\ substitution]}{[\#\ phones\ in\ target\ word]} < \beta \quad (2)$$

where the longest common sub-sequence (LCS) is found as the longest sub-sequence present in both the target word and the substitution, for example the LCS for the sequences "*C A T S*" and "*C A R T*" is "*C A T*". A substitution is excluded when more than α of its phones are in the longest common sub-sequence, as expressed in (1), *and* when the number of phones in the substitution is less than β of the number of phones in the target word as in (2).

The forced-alignment type language model with optional substitution words added can be seen in Figure 1. The relationship between the transition probability of the target word P_T and the substitution words P_S determines the number of miscues detected, the larger P_S gets as compared to P_T the more miscues are detected at the cost of more rejections. We detect a miscue when the alternative substitution word is recognized instead of the target word.

We only add real word substitutions to the language model since nonsense miscues have not been phonetically transcribed.

3.3. Goodness of pronunciation

The goodness of pronunciation score was presented by [8] for use in pronunciation assessment of language learners. The score is implemented using a forced-alignment and a free phone loop speech recognizer and can be expressed as:

$$GOP(n) = \left|LLF(n) - LLP(n)\right| \quad (3)$$

where $LLF(n)$ is the acoustic log likelihood of frame n from the forced-alignment decoder, and $LLP(n)$ is the acoustic log likelihood of frame n from the free phone loop decoder. We then calculate the time-normalized GOP score for each forced-aligned phone, and mark words having at least one phone with a GOP score higher than a predefined threshold as miscues. The value of the GOP threshold determines the number of miscues

detected, the higher the threshold, the lower the number of detected miscues.

The GOP score can be used for miscue detection when it's combined with the forced-alignment language model and the fact that we only mark words as miscues that have not been pronounced correctly at least once (see Section 5); we actually use this setup to look for correctly read words instead of miscues.

4. Corpus of dyslexic read speech

The speech corpus used in the experiments consists of read speech from 75 Danish adult dyslexic readers at undergraduate educations. The same target text was read by all readers and consists of 10 sentences and contains 221 words. Of the 193 minutes of audio, 117 minutes contain speech. Of the 17665 transcribed speech events (correctly read words and miscues) 3127 are marked miscues – giving a miscue rate of 18%. The number of real word substitution miscues in the entire corpus is 378 (12%).

The speech data has been collected by academic reading tutors who specialize in tutoring dyslexics. The recordings have been made in quiet office environments using laptop computers and table top microphones. The sample frequency is 44.1 kHz at a resolution of 16 bits/sample. The reader was given a piece of paper with the target text and instructed to read it out loud. The reader then read the text without any interventions by the reading tutor.

The recordings have been transcribed on the word level and each incorrectly read word has been labeled with the miscue types presented in Section 2 except for long pauses between words. This means that some of the miscues will be very hard to detect e.g. elongation of consonant or vowel sounds, some mispronunciations, and unnatural prosody.

5. Evaluation

We choose to mark target text words as miscues only if they are not read correctly at least once. This means that we ignore all sentence level miscues except for word omissions. Word level miscues that are corrected by the reader are also ignored.

The corpus of dyslexic read speech is split into a training set of 37 speakers, a development set of 19 speakers, and a test set of 19 speakers. We extract the substitution miscues from the training set and use the development set for tuning the automatic speech recognizer. We use the test set for evaluating miscue detection accuracies.

We base our miscue detection system on the Sphinx-4 speech recognizer and train the acoustic models using SphinxTrain – both are part of the CMU Sphinx group's open source speech recognition engines [9]. The hidden Markov models are the same as used in [4]: context-dependent, tied-state, tri-state left-to-right hidden Markov models with 16 Gaussians per mixture.

6. Results and discussion

The results are presented as miscue detection rates (MDR) and false rejection rates (FRR). Miscue detection rate is calculated as the number of correctly rejected words divided by the number of actual reading miscues. False rejection rate is calculated as the number of falsely rejected words divided by the number of correctly read words.

We have set α and β in Equation 2 to 0.5 for all experiments. At this level the included substitutions cover $64/122 = 53\%$ of the substitution miscues seen in the test set. Figure 2 shows MDR vs. FRR when varying the GOP threshold for a number of different substitution penalties. The circles indicate infinite GOP thresholds at the substitution probabilities listed below the circles. The curve that starts at $(0, 0)$ shows the performance of the GOP score alone, i.e. the substitution penalty has been set infinitely high.

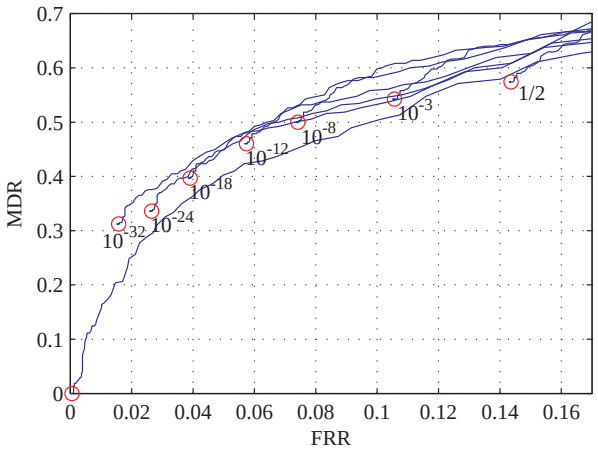

Figure 2: *MDR vs. FRR. The circles indicate infinite GOP thresholds*

We see that both the GOP score and the substitution approach do worse by themselves, than the combination of the two. The effect of combining the language model with the GOP score is weakest at miscue detection rates around 45%. However, at a miscue detection rate of 57% using only the specialized language model results in a false rejection rate (FRR) of 14%, using only the GOP score results in a FRR of 13%, whereas a combination of the two results in a FRR of 9%. Similarly at a miscue detection rate of 34% (1063 miscues) the FRR using only the specialized LM is 2.6% (459 correctly read words), for the GOP score it's 3.4% (601 correctly read words), whereas a combination has a FRR of only 1.8% (318 correctly read words) – a relative reduction of 31% when compared to the FRR of the specialized LM approach.

One example of where the two detection methods supplement each other well is for an utterance with the target text "...eller hørespilsdramaturgerne og ... de stadig sjældnere forskningsstipendier.". Here the reader mispronounced and split the word "høresplisdramaturgerne" in two and read the word "sjældnere" as "sjældne" (inflection mistake). The first word is marked as a miscue by the GOP score but is not caught by the LM approach, since this non-real word is not in the LM. The second word is marked as a miscue by the LM approach since the inflection mistake is modeled in the LM; but is not caught by the GOP score, since the pronunciation is very close to the target word. In this example, each of the two detection methods only detects a single miscue but when combined they detect both.

7. Conclusions

We have presented a miscue detection method that combines a specialized language model, augmented with real word substitutions, and the goodness of pronunciation (GOP) score. Experiments show that this combination improves miscue detec-

tion performance. At a miscue detection rate of 34% the false rejection rate using only the specialized LM is 2.6%, for the GOP score it's 3.4%, whereas a combination of the two miscue detection methods has a false rejection rate of only 1.8%.

For future work we would like to exploit more of the miscue statistics of the corpus of dyslexic read speech in order to improve detection performance at the same or lower false alarm rates.

8. Acknowledgments

We would like to thank Rådgivnings- og støttecentret (The Advice and Support Centre) [10], Aarhus University, for providing the acoustic material.

Most of the work presented in this paper has been sponsored by Oticon Fonden (The Oticon Foundation) [11].

9. References

[1] Duchateau, J., Kong, Y. O., Cleuren, L., Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Werner Verhelst, W., and hamme, H. V., "Developing a Reading Tutor: Design and Evaluation of Dedicated Speech Recognition and Synthesis Modules", Speech Communication, volume 51, No. 10, October 2009, pp. 985–994.

[2] Andreas Hagen, Bryan Pellom, Ronald Cole, "Highly accurate childrens speech recognition for interactive reading tutors using subword units", Speech Communication, vol. 49, no. 12, December 2007, pp. 861–873.

[3] Yik-Cheung Tam, Jack Mostow, Joseph Beck, Satanjeev Banerjee, "Training a Confidence Measure for a Reading Tutor that Listens", Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, September 2003, pp. 3161–3164.

[4] Rasmussen, M. H., Tan, Z.-H., Lindberg, B., and Jensen, S. H., "A System for Detecting Miscues in Dyslexic Read Speech", European Conference on Speech Communication and Technology (Interspeech), Brighton, U.K., 2009, pp. 1467–1470.

[5] Mostow, J., Beck, J., Winter, S. V., Wang, S., and Tobin, B., "Predicting oral reading miscues", Seventh International Conference on Spoken Language Processing (ICSLP-02), Denver, CO, September 2002, pp. 1221–1224.

[6] Mostow, J., Roth, S., Hauptmann, A. G., and Kane, M., "A Prototype Reading Coach that Listens", Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), American Association for Artificial Intelligence, Seattle, WA, August 1994, pp. 785–792.

[7] Banerjee, S., Beck, J., and Mostow, J., "Evaluating the Effect of Predicting Oral Reading Miscues", Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, September 2003, pp. 3165–3168.

[8] Witt, S. M., "Use of Speech Recognition in Computer-assisted Language Learning", Ph.D. thesis, University of Cambridge, 1999.

[9] Speech at CMU, "The CMU Sphinx Group Open Source Speech Recognition Engines", Carnegie Mellon University, Online: http://cmusphinx.sourceforge.net, accessed on the 18th of April 2011.

[10] Rådgivnings- og støttecentret, Online: http://www.dpu.dk/rsc/, accessed on June 30, 2011.

[11] Oticon Fonden, Online: http://www.oticonfonden.dk, accessed on March 29, 2011.