INTERSPEECH 2009 BRIGHTON

A System for Detecting Miscues in Dyslexic Read Speech

Morten Højfeldt Rasmussen, Zheng-Hua Tan, Børge Lindberg and Søren Holdt Jensen

Multimedia Information and Signal Processing, Department of Electronic Systems,
Aalborg University, Denmark
{mr,zt,bli,shj}@es.aau.dk

Abstract

While miscue detection in general is a well explored research field little attention has so far been paid to miscue detection in dyslexic read speech. This domain differs substantially from the domains that are commonly researched, as for example dyslexic read speech includes frequent regressions and long pauses between words. A system detecting miscues in dyslexic read speech is presented. It includes an ASR component employing a forced-alignment like grammar adjusted for dyslexic input and uses the GOP score and phone duration to accept or reject the read words. Experimental results show that the system detects miscues at a false alarm rate of 5.3% and a miscue detection rate of 40.1%. These results are worse than current state of the art reading tutors perhaps indicating that dyslexic read speech is a challenge to handle.

Index Terms: miscue detection, reading tutor, dyslexia, speech recognition, confidence score

1. Introduction

The advances in automatic speech recognition (ASR) significantly boost the research and development of automatic reading tutors, especially for children and second language learners. Automatic reading tutors for dyslexics, however, are rarely investigated. Such investigations are important as dyslexia is a rather common learning disability; the number of dyslexics in Denmark for example is around 2-5% of the population [1]. Human tutors are very costly and as a result, many students cannot get the help they need.

A method for teaching the dyslexics to become better readers is that of the book-and-tape approach [1]. By this method the dyslexic is taught to read in the following way: first a text is presented to him, then a prerecording of the read text is played back once after which the dyslexic reads the same passage; if the dyslexic is having problems reading the text a human tutor will help him. This approach has been automated by [1] using the miscue detection system presented in this paper since automating the book-and-tape approach needs automatic miscue detection in order to be able to provide feedback to the dyslexic.

The field of miscue detection and confidence scoring is large and a lot of research has been conducted within the area as made evident by e.g. the survey [2]. This is also true for the specific area of miscue detection for reading tutors. Miscue detectors for second language learning, e.g. [3], and reading tutors for children, e.g. [4], [5] and [6], have been created. While miscue detection in general is a well explored research field little attention has so far been paid to miscue detection in dyslexic read speech. This domain differs substantially from the domains that are commonly researched, as for example dyslexic read speech includes frequent regressions and long pauses between words [1].

This paper presents a system that automatically spots reading miscues in sentences read by dyslexic persons speaking Danish. The core of the system includes an automatic speech recognition system using a forced-alignment grammar and a number of miscue detection methods. A real-time system is built and the accuracy is determined.

The rest of the paper is structured as follows: in section 2 the methods for spotting reading miscues in the presented system are described, in section 3 the databases and the ASR used are presented, in section 4 the performance of the system is evaluated and compared to state-of-the-art reading tutors for children and in section 5 the results are discussed and conclusions given.

2. Miscue detection in dyslexic read speech

Dyslexic read speech differs significantly from a normal reading person's read speech in a number of ways. Some of the miscues encountered in dyslexic read speech are [1]:

- Regressions;
- Filled pauses;
- Long pauses between words;
- Word skipping;
- Word truncation.

Furthermore, the speech can also be hesitant and soft when the reader is insecure.

Due to the nature of dyslexic read speech it is debatable what is meant by miscues and correctly read words. In this paper correctly read words are defined as words correctly read at least once. Accordingly all regression and insertion miscues are therefore disregarded.

Three approaches to detect reading miscues are implemented in the presented system. The three methods are based on forced-alignment, likelihood and phone duration.

2.1. Forced-alignment method

To be able to detect reading miscues the language model used in the system are based on the text the reader is supposed to read. It is implemented as a finite state forced-alignment grammar with transitions from word $n-1$ to n in the text. However, in order to allow for dyslexic reading patterns, a forced-alignment grammar in the strictest sense cannot be used. The grammar used in the presented system is shown in Figure 1.

Each arc in the grammar contains filler models (FMs) that are allowed to be skipped. The FMs are comprised of a silence model, a model of filled pauses, a model of speaker noise (coughs, lip smacks, etc.) and a garbage model trained on all phones the training database. One of the functions of the inserted FMs are to handle regressions and word insertions. Used

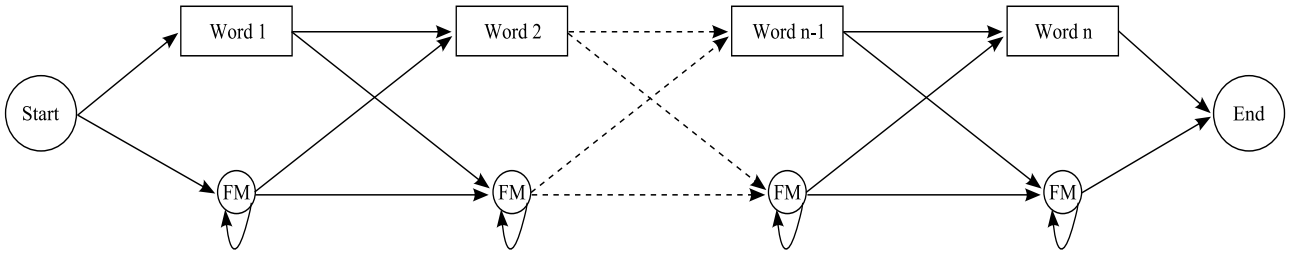

Figure 1: *The forced-alignment like grammar used in the system. [7]*

in this way, the (most) correct instance of the word/sentence being repeated will be aligned to the word/sentence of the forced-alignment grammar and the regressions and insertions will be aligned to the garbage model, effectively ignoring the added words. The inserted FMs also model the filled pauses and the long silences between words.

To model skipping of words, an optional path going forward two or more words through FMs is added. To ensure that the decoder does not favor the garbage loop and therefore aligns the loop to the whole utterance, the probability of skipping a word is set much lower than the probability of not skipping a word. Empirical observations indicate, that the probability of skipping a word should be around 10^{-6} of the probability of not skipping a word. This value corresponds to the probability of seeing a word skip in dyslexic speech and taking into account the general acoustic fit of the GM compared to that of the competing word.

2.2. Likelihood method

The likelihood based method works by accepting or rejecting a word based on an estimate of the maximum a posteriori (MAP) likelihood called the goodness of pronunciation (GOP) score by [3]. Using the output of a forced-alignment decoder and a free phone loop decoder, the GOP score for the n^{th} speech frame is calculated as:

$$GOP(n) = \left|LLF(n) - LLP(n)\right| \tag{1}$$

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. The grammar used by the phone loop decoder is shown in Figure 2, where p1, p2, . . . , pn are monophones.

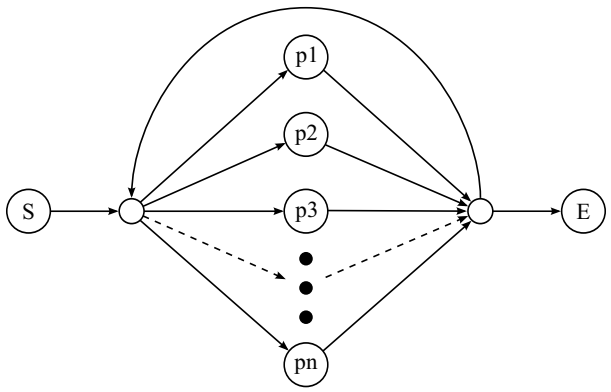

Figure 2: *The phone loop grammar used in the system.*

The GOP score will be calculated on both the word level and the phone level. The word level GOP score is calculated as the average frame GOP score of a word and the phone level GOP score as the average frame GOP score of a phone. A word can be classified as a reading error if either the word level GOP score or the highest phone level GOP score of a word are above predefined thresholds. The boundaries used when calculating the average GOP score of the words and phones are determined by the output from the forced-alignment decoder. The threshold determines how "strict" the system is; that is, if the threshold is high the system labels fewer words as reading errors and vice-versa.

2.3. Phone duration method

The phone duration based method works by accepting words with phones that have durations close to the expected durations and rejecting words with one or more phones that are either considered too short or too long. A phone realization is considered too short if:

$$\frac{D_{p_f}}{D_{p_n}} < T_{p_s} \tag{2}$$

where D_{p_f} is the duration of the force-aligned phone realization, D_{p_n} is the expected phone length and T_{p_s} is a predefined "short phone" duration threshold. Similarly a phone is considered too long if:

$$\frac{D_{p_f}}{D_{p_n}} > T_{p_l} \tag{3}$$

where T_{p_l} is a predefined "long phone" duration threshold.

Since the hidden Markov models (HMMs) are left-to-right models, and skips are not allowed, the expected phone length (D_{p_n}) can be estimated from the probabilities in the HMM transition matrix as:

$$D_{p_n} = D_f \sum_{j=1}^{N_s} \frac{\log(a_j)}{\log(b_j)} \tag{4}$$

where D_f is the length of a frame shift (in this case D_f is 10 ms), N_s is the number of states in a HMM, a_j is the probability of going forward one state from state j to $j+1$ and b_j is the probability of remaining in state j.

3. Data and ASR

The database used for training the acoustic models is collected as part of a project to make a command-and-control application called "Indtal" in Danish [8]. The database contains 30 hours of speech from 450 adult native Danish speakers from various parts of the country. Since the database is small, all the material is used for training the acoustic models, including non-sentence utterances like application commands. The recordings have been made in quiet environments using the same type of

close-talking microphone at a sampling rate of 16 kHz in 16 bit resolution.

The test database used in all of the experiments contains 7600 words (1350 of which are miscues) in 100 minutes of speech and silence from 8 adult dyslexic persons (see [1] for more details). The recordings have been made in the same way as the recordings of the Indtal database and transcribed according to the SpeechDat(II) [9] transcription standard with the extension of the following layers:

- Error types;
- Intended words;
- Time-related annotation;

For what concerns the error types layer, word substitutions, insertions and splittings are annotated. At the time-related layer, pauses and jumps (both forward and backward) and annotated. Each word has thus been marked as either correctly or incorrectly read. Each speaker is processed separately; resulting in 8 decoder runs.

The automatic speech recognizer used in the system is the Sphinx-4 recognizer and the acoustic models are trained using SphinxTrain – both are part of the CMU Sphinx group's open source speech recognition engines [10]. All HMMs are context-dependent, tied-state, tri-state left-to-right HMM with 16 Gaussians per mixture. Skipping of states on the HMM level is not allowed. The features extracted by the front-end are the 13 first MFCC plus first and second order derivatives. A finite state grammar (FSG) is created for each speaker/text based on the prompt text.

4. Experiments

In order to compare the performance of the four detectors (GOP on word and phone level plus short and long phone duration), the accuracy of each detector is calculated as the area under its receiver operating characteristic (ROC) curve. The ROC curves are created by plotting the miscue detection rate (MDR) as a function of the false alarm rate (FAR) for varying thresholds. MDR is calculated as the number of times the method correctly rejects a word normalized by the number of miscues made by the reader. FAR is calculated as the number of times the method wrongly rejects a word normalized by the number of correctly read words.

A simple method for combining the four detectors will be tried out. The system will mark a word as being a miscue if any of the four detectors detects a miscue. All combinations of thresholds at regular intervals within the space of the prompted words will be tried out. With the selected resolution this gives approximately 150k different threshold combinations.

The resulting ROC curves for the four detectors and the result of combining them are shown in Figure 3 and the calculated accuracy of the setups is shown in Table 1. The light blue area beneath the curve describing the performance of the phone level GOP score is comprised of the dots describing the performance of the 150k different combinations of the detectors. Since the forced-alignment grammar used allows word skips, the ROC curve will not start in (0,0). Because of this, the lowest FAR value achievable in this system is 5.3%, and the largest possible accuracy is 94.7%.

As the system is intended to be used as a remedial tool a low FAR is desired. At the lowest FAR value of 5.3% the MDR is at 40.1%. At this point the thresholds are set high enough (or in the case of the "short phone" duration; low enough) to mark all words as being correctly read and all marked miscues are therefore attributable to the forced-alignment grammar.

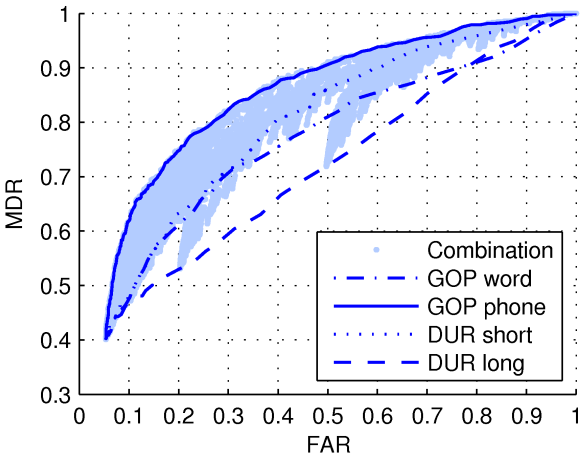

Figure 3: *Performance of the four features separately and combined.*

Table 1: *Miscue detection accuracies for different detectors.*

Detector	Accuracy
GOP_{word}	74.4%
GOP_{phone}	82.9 %
DUR_{short}	77.4%
DUR_{long}	69 .6%

As mentioned in the introduction, the presented system has been implemented in an automated version of the book-and-tape approach. The resulting system has been tested in a field trial [1]. The result of the field trial shows that the dyslexic users are positive towards such a system, even with the imperfect miscue detection.

From the ROC curves and the accuracies it is evident that the best performing method is the phone level GOP score. Even the combination of the methods does not perform much better than the phone level GOP score alone; at any given FAR the best combination is at best only a few detections better.

It can also be seen that all the detection methods except for the "long phone" duration based are better than randomly marking miscues (the line of no discrimination starts at $(5.3\%, 40.1\%)$ and ends at $(100\%, 100\%)$).

Other reading tutor systems (but for children) achieve better results; for example [5] achieves a FAR and MDR of around 3% and 58% respectively and [4] achieves a very low FAR of 0.5% at a MDR of 80.0%. The performance difference seems large, but since the training data and application domains are different, it is hard to determine if the performance difference is because of the system setup or that dyslexic read speech is more difficult to handle for an ASR system than normal reading persons, children and adults, read speech.

On average the setups run at 1.3 times real time on a 2.8 GHz Pentium 4 computer.

5. Conclusion

In this paper, a system for detecting miscues in Danish dyslexic read speech has been presented. The system uses methods that

build on the GOP scoring algorithm and phone duration and uses a forced-alignment grammar tuned for dyslexic input for spotting the miscues. Experimental results show that the system works for the task of detecting miscues in dyslexic read speech. However, the performance of the system applied on dyslexic read speech is low compared to reading tutors for different domains which might be an indication of how difficult it is to handle dyslexic read speech. This will be explored further in future work.

6. Acknowledgements

PhD student Morten Højfeldt Rasmussen is supported by the Oticon Foundation [11]. The scientific responsibility is assumed by the authors. The authors would also like to thank our colleague Dr. Jakob Schou Pedersen for collecting and annotating the database containing dyslexic read speech used for the experiments presented in this paper.

7. References

[1] Pedersen, J. S., "User Centred Design Of a Multimodal Reading Training System for Dyslexics", Ph.D. thesis, Aalborg University, 2009.

[2] Hui Jiang, "Confidence measures for speech recognition: A survey", Speech Communication, vol. 45, no. 4, 455-470, April 2005.

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

[4] Jacques Duchateau, Mari Wigham, Kris Demuynck, Hugo Van hamme, "A Flexible Recogniser Architecture in a Reading Tutor for Children", In Proc. ITRW on Speech Recognition and Intrinsic Variation, Toulouse, France, 2006.

[5] Yik-Cheung Tam, Jack Mostow, Joseph Beck, Satanjeev Banerjee, "Training a Confidence Measure for a Reading Tutor that Listens", Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, 2003.

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

[7] Pedersen, J. S., Rasmussen, M. H., Lindberg, B., "ASR and Dyslexic Input", technical report, ISSN: 0908-1224, Aalborg University, 2007.

[8] Brøndsted, T., Aaskoven, E., "Voice-controlled Internet Browsing for Motor-handicapped Users: Design and Implementation Issues", in procedings for the 9th European Conference on Speech Communication and Technology (Interspeech), Lisboa, Portugal, 2006.

[9] Lindberg, B., "Speechdat, Danish FDB 4000 speaker database for the fixed telephone network", EU-Project SpeechDat, LE2-4001, 1998.

[10] Speech at CMU, "The CMU Sphinx Group Open Source Speech Recognition Engines", Carnegie Mellon University, Online: http://cmusphinx.sourceforge.net, accessed on the 14th of April 2008.

[11] The Oticon Foundation, "Oticon Fonden", Online: http://www.oticonfonden.dk, accessed on the 18th of June 2009.