Developing an Automatic Assessment Tool for Children's Oral Reading

Leen Cleuren, Jacques Duchateau, Alain Sips, Pol Ghesquière and Hugo Van hamme

Katholieke Universiteit Leuven, Belgium

E-mail: Leen.Cleuren@ped.kuleuven.be, Jacques.Duchateau@esat.kuleuven.be

Abstract

Automation of oral reading assessment and of feedback in a reading tutor is a very challenging task. This paper describes our research aiming at developing such automated systems. First topic is the recording and annotation of CHOREC, the Flemish database of children's oral reading we develop in order to characterize oral reading processes statistically. Next, we propose a classification of both oral reading strategies and errors, which provides the basis of the envisaged assessment and feedback. Finally, experimental results show that our two-layered recognition system is able to provide high reading miscue detection rates, while only few correctly read words are erroneously tagged as miscue.

Index Terms: reading assessment, database annotation, speech technology, education.

1. Introduction

In Flanders, primary school children's progress is regularly assessed in order to detect early learning difficulties such as reading disabilities (RD). With respect to the child's reading development, much attention is paid to the assessment of word decoding skills as these are crucial for adequate reading. Whenever word decoding problems arise, early, regular and adequate intervention is highly needed in order to prevent the child from dropping further and further behind his or her classmates, not only with respect to his or her reading development but also with respect to appropriate functioning in other classes since text usually plays an important role in all of them.

To detect RD in children, various valuable paper-and-pencil diagnostic instruments are being used at the moment [1, 2, 3]. Although the quality of these instruments isn't questioned, it is clear that the administration of these tests is very time consuming. Another disadvantage of this way of assessment has to do with the fact that, although usually performed by experienced examiners, the evaluation suffers from an examiner bias.

By improving speech recognition technology in order to make it suitable as a supporting technology for an automated reading tutor, the SPACE project[1] wants to account for both disadvantages of traditional word decoding assessment. By means of a speech recognizer, the oral reading assessment can become automated and objective measures of the child's reading performance can be obtained. Moreover, if a reading tutor can really listen to a child reading aloud, and if natural, highly intelligible and phonetically correct speech can be synthesized, it is also possible to build interactive tools so that a reading tutor can act as a fluent reader model and is able to provide immediate feedback to the learning child.

[1] SPACE: SPeech Algorithms for Clinical and Educational applications. Home page: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

In spite of the availability of the text being read, automatic speech recognition within the context of reading assessment and instruction is a very challenging task due to reading-related and child-related developmental processes. For example, due to a child's variable maturation, articulatory competencies differ and due to variable word decoding skills, oral reading of novice readers or readers with RD can be fraught with oral reading errors.

In order to improve the speech recognizer's ability to accurately detect reading errors, it is necessary to characterize reading behavior statistically and to provide the recognizer with a model that contains information on the nature and prevalence of likely reading errors. To achieve this goal, the CHOREC (CHildren's Oral REading Corpus) database of recorded, transcribed and annotated children's oral reading and oral reading errors is being developed.

This paper is organized as follows. Section 2 gives a brief description of the CHOREC database. The way the recorded reading sessions are transcribed and annotated, is explained in section 3. In the next section, two tiers in the annotation are highlighted: the oral reading strategy tier and the oral reading error tier. Section 5 deals with the evaluation of our baseline reading miscue detector. Finally, in section 6, some conclusions and ideas for future work are given.

2. A database of children's oral reading

The CHOREC database contains reading sessions of Dutch-speaking elementary school children (grade 1-4). Until now approximately 300 children were recorded; in the near future the database will be extended by including children with known reading disabilities.

For every child, a newly developed computerized reading test battery is administered, containing a real word reading test (RWRT), a pseudoword reading test (PWRT) and a story reading test (SRT). Both the RWRT and the PWRT contain three lists of respectively 40 1-syllable, 40 2-syllable and 40 3- or 4-syllable real words or pseudowords. The SRT consists of 9 graded text stories (vocabulary of 538 distinct word forms), ranging from AVI 1 to AVI 9[2] in difficulty and 103 to 223 words in length. Real words and pseudowords are presented individually; text stories are presented paragraph by paragraph.

Children are selected from grade 1 to 4, meaning that, on average, they are 6 to 10 years old. Each child reads minimally one and maximally three real word lists and pseudoword lists, and minimally one and maximally four text stories depending on that particular child's reading level. Relevant information about each child

[2] In the Netherlands and Flanders, the AVI-index is used to distinguish between texts of different technical difficulty level. The index is largely based on the reading index A [4] which takes into account word, sentence and text features.

Expected	Els zoekt haar schoen onder het bed. [Els looks for her shoe under the bed.]
Observed	Als zoekt haar sch... schoen onder bed. [Als looks for her sh... shoe under bed.]

Tier 1	Els	zoekt	haar	schoen	onder	het	bed.
Tier 2	Els	zoekt	haar	schoen	onder	het	bed.
Tier 3	*	zoekt	haar	* schoen	onder		bed
Tier 4	Als	zukt	har	sx sxun	Ond@r		bEt
Tier 5	df	dg	dg	sg	dg	O	dg
Tier 6	4					36	

Figure 1: *Transcription and annotation of a read sentence by means of 'Praat'(Tier 1-6). [Tier 5 - df: incorrect decoding within the first trial; dg: correct decoding within the first trial; sg: partially spelling out before correctly synthesizing; O: omission. Tier 6 - code 4: vowel substitution; code 36: omission of a word.]*

(sex, age, grade, hight, curriculum, place of birth, place of residence, mother tongue, reading level, reading method, presence of reading disabilities) is carefully gathered. Two-channel recordings are made, using a microphone pinned on the child's shirt and a table top microphone.

3. Transcription and annotation of oral reading sessions

The recorded reading sessions are transcribed and annotated manually by means of a customized version of 'Praat' (http://www.Praat.org/), a free-source tool for speech analysis and synthesis. This tool provides the possibility to attach a text-grid (containing different tiers) to the speech sound. As such, each tier provides another layer of descriptive information about the speech sound that particular tier is attached to.

The 8 annotation tiers used in the CHOREC database include both information on utterances directly resulting from the child's efforts to read what is presented on the computer screen as well as information on background noise and reading task-related and -unrelated unforeseen utterances made by both the child and the examiner.

Each tier originally contains segment boundaries corresponding to the timing of new stimulus presentation. In case of the RWRT and PWRT, boundaries indicate the presentation of a new real word or pseudoword; in case of the SRT, they indicate the presentation of a new paragraph. An example of the information available in the proposed tiers in CHOREC is given in figure 1.

Tier 1. In order to carefully describe and annotate oral reading recordings, it is often advisable to move or remove existing boundaries or insert new ones whenever the child hesitates or makes a reading error. This implicates however that information about the exact timing of reading stimulus presentation is lost. Because latency and production times can be of great importance for situating readers within a developmental model of reading [5] or for discriminating readers with reading disabilities from those without, tier 1 is used for preserving the original boundaries (and the original reading task) corresponding to the exact timings of word and paragraph presentation.

Tier 2. As in tier 1, tier 2 doesn't allow for making changes with respect to the orthographic transcription of the original reading task. Segment boundaries, however, can be manually moved or removed and newly inserted when necessary for annotation. This means that this tier allows for the insertion of new segments whenever needed. Segment boundaries in tiers 3, 4, 5, and 6, will be linked to these in tier 2.

Tier 3 and 4. As is the case in tier 1 and 2, tier 3 initially contains the orthographic transcription of the original reading task. Tier 4 provides us with the phonetic transcription of what is actually read: originally, this tier consists of a concatenation of lexicalized phonetic transcriptions, including all possible correct pronunciations but ignoring cross word assimilation. If the reading task is read fluently without any errors or hesitations, tier 3 and 4 remain unchanged. However, whenever the child struggles, makes reading errors (also including omissions, real word-substitutions and insertions) in a particular word or inserts words or interjections unrelated to the printed reading task, manual adjustments to this word [3] have to be made in tier 3 and 4 in order to capture each of these reading attempts and errors. Appearance of assimilation between words or dialect influences within a word is only captured whenever changes have to be made to tier 3 and/or tier 4.

Tier 5 and 6. Tier 5 provides the possibility to classify oral reading behavior used to arrive at the correct or incorrect reading of words, possibly revealing different reading strategies. By letting the reading tutor keep track of a child's reading behavior preferences, useful information is gathered in order to distinguish readers with RD from those without. Tier 6 is used for classifying the type of oral reading error occurring at a particular word during reading. To do so, an oral reading error classification based on surface level features, is used. These tiers are discussed in section 4 in more detail.

Tier 7 and 8. Tier 7 is used to orthographically transcribe possible extra utterances by the examiner. Segment boundaries are placed so as to surround the utterance as closely as possible. In tier 8, boundaries are put to catch background noise.

[3] For convenience reasons, 'word' is used for real words, pseudowords as well as words coming from a text story.

4. Classification of reading strategies and errors

4.1. Reading strategies

The reading of children facing reading disabilities (RD) is generally characterized by either a compensatory guessing or (phoneme-by-phoneme/syllable-by-syllable) spelling strategy. Although both strategies can be useful for any beginning reader, children with RD use them too excessively. By using the strategy of guessing, one is able to keep up a reasonable reading rate but doesn't really decode the printed word anymore whereas a reader using a spelling strategy, puts so much effort in decoding that the reading is slow and laborious.

Strategies for arriving at the correct reading of words, besides correct decoding within the first trial, are:

- repeating a correctly decoded word once or more;
- partially or completely correctly or incorrectly spelling out a word before correctly synthesizing it (revealing a spelling strategy);
- incorrectly decoding a word in the first trial but reading it correctly in the final trial (possibly revealing a guessing strategy).

Strategies that do not result in the correct reading of words, besides incorrect decoding within the first trial (possibly revealing a guessing strategy), are:

- partially or completely correctly or incorrectly spelling out and incorrectly synthesizing a word or not synthesizing it at all (revealing a spelling strategy);
- correctly decoding a word within the first trial but then 'correcting' it wrongly in a second trial (possibly revealing a guessing strategy);
- mistakenly omitting or inserting a word;
- asking for a partial or complete prompt of a word before carrying on reading it.

4.2. Reading errors

The ultimate goal of researchers studying reading disabilities (RD) is to obtain an accurate characterization of the phenotypic performance pattern of children with RD. Therefore, the analysis of reading errors has also long attracted the attention of reading researchers. Many have seen it as a useful way to increase our understanding of the reading process and as a basis for making decisions about classroom instruction [6, 7].

At the same time, within the scope of the SPACE project, reading error analysis is needed in order to identify a set of likely reading errors that eventually can be modeled in the speech recognizer. If an automated reading tutor would listen for every possible phoneme sequence in place of a correct word, this would result in too many recognition errors, given the limited accuracy of current speech recognizers [8].

Reading errors can occur at four different levels: at the level of the paragraph, the sentence, the word, or the grapheme. At the paragraph level, errors include omissions or repetitions of a whole line or sentence. At the sentence level, a part of a sentence can be omitted or repeated; a word can be mistakenly inserted; a word can change places with another word in the sentence; or a word can be substituted for a word being a synonym of or semantically related to that word.

At the level of the word, reading errors are generally characterized by errors directly resulting from an incorrect compensatory decoding strategy. A spelling strategy can result in letter-by-letter or syllable-by-syllable reading without (correctly) synthesizing; and a guessing strategy can result in the substitution of a word or pseudoword by another word or pseudoword respectively, or the substitution of a pseudoword by a real word.

Four main categories of reading errors can be distinguished at the grapheme level: sequential errors, substitution errors, deletion errors, and insertion errors. Sequential errors include changing letter order of adjacent and nonadjacent letters. Substitution errors include vowel substitutions, consonant or consonant cluster substitutions, and vowel-consonant substitutions. Deletion and insertion errors are characterized by vowel deletions resp. insertions, consonant or consonant cluster deletions resp. insertions, and syllable deletions resp. insertions.

5. Automatic reading miscue detection

5.1. Baseline system description

The speech recognition system that was used for the automatic reading miscue detection has a two-layered architecture. In the first layer, a generic phoneme recognizer is used to produce a phoneme lattice. This phoneme recognizer involves a general N-gram phoneme sequence model for the language at hand; it does not include information on the words or sentences that should be read by the child. The second layer contains all task specific information. A finite state transducer (FST) models the words in the sentence to be read but also includes solutions for both expected reading miscues and unexpected events and disfluencies. Based on this phoneme level FST, the best path through the phoneme lattice is looked for and is returned as recognition result.

The baseline FST we currently implement, models the sequence of words that should be read and the acceptable pronunciations of those words. In addition, two expected reading miscues are included. First, arcs are added from every node in each word model to the start of the word in order to cope with partially read words when the child restarts the word to try again. Second, extra arcs are added in the sentence model to allow for skipping and repeating words. However, we can never suppose that a child will only produce expected reading miscues. Therefore, a garbage model is added both between the expected words and in parallel with each expected word. The garbage model is a phoneme loop based on the N-gram phoneme sequence model. More information about the proposed reading miscue detection system can be found in [9].

In the future, we will improve and refine the modeling of the expected miscues based on the classification of reading strategies and errors described in section 4, and on the statistical information gathered from the CHOREC database.

5.2. System evaluation

The first experiments to evaluate the miscue detection system are not based on the CHOREC database as the annotations were not available at the time. Instead we use a read speech database in Dutch which contains recordings of children, aged between 5 and 11, reading sentences. The children are divided into training speakers and test speakers. The training speakers are used for the acoustic modeling which includes a VTLN (Vocal Tract Length Normalization) method that does not introduce latency in recog-

	age 6	age 9 to 11
miscue detection rate	71.7%	80.0%
false alarm rate	0.5%	0.5%

Table 1: Miscue detection and false alarm rates depending on the child's age

nition (this is for instance needed for tracking in a reading tutor). More details about this, and other experimental results, can also be found in [9].

In order to allow the use of the read speech database for an experiment on miscue detection, the orthographic transcriptions in the test set sentences were transformed into strings of correctly read words and miscue markers. Also, the result of our recognition system (described above) is turned into such a string: each time an acceptable phonetic transcription of a word is found in the phoneme lattice, this word is put in the result string, while each time the system has to follow any of the paths in the FST that is meant for the expected or unexpected miscues, a miscue marker is put in the result string. By aligning both strings, we can count the number of miscues that are (or are not) detected by the recognition system, as well as the number of false alarms, when a correctly read word is tagged as miscue.

In table 1, we analyze the system by its miscue detection rate (the number of miscues correctly detected divided by the total number of miscues the child made) and its false alarm rate (the number of words erroneously flagged as read incorrectly divided by the total number of words the child read correctly).

Results are given for 2 distinct age groups: 6 year olds and children between 9 and 11 years old. We can see that miscue detection is more difficult for younger children. The reason is that the acoustic decoding contains more errors for the youngest children: the generic phoneme recognizer (which is the first layer in the proposed system) results in a phoneme error rate[4] of 33.0% for 6 year old children and 21.4% for 9 year old children.

The miscue detection rates in the table are well comparable to the values for other state-of-the-art systems [10, 11], which recently reported miscue detection rates of 56% and 67% respectively. But these systems result in a false alarm rate of about 3% and they seem to be incapable of producing lower false alarm rates at a reasonable miscue detection rate. It should be noted however, that comparing the systems is rather tricky, as none of the systems uses the same test data, e.g. the age of the children is critical and may differ, and the language involved may also matter.

6. Conclusions

A speech recognizer is only able to accurately recognize reading errors if we can model the nature and prevalence of likely reading errors. In order to characterize this children's reading behavior statistically, we are developing CHOREC, a database of recorded, carefully transcribed and annotated children's reading and reading errors. We also investigate a baseline reading error model in a reading miscue detector and find promising results.

Future research will include the recording and annotation of reading disabled children's reading sessions, the analysis and modeling of their reading strategies and errors, and the upgrade of the current miscue detection system to an automatic miscue classification system. The latter will be part of the envisaged automated reading assessment and reading tutor.

[4] This is the sum of substitutions, insertions and deletions in the recognized phoneme string with respect to the number of phonemes in the reference string.

7. Acknowledgment

The research in this paper was supported by the IWT project SPACE (sbo/040102): SPeech Algorithms for Clinical and Educational applications, home page: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

8. References

[1] J. Visser, A. van Laarhoven, and A. ter Beek, *AVI-toetspakket. Handleiding.*, KPC Groep, 's Hertogenbosch, 1996.

[2] L. Verhoeven, *Drie-Minuten-Toets. Handleiding.*, Cito, Arnhem, 1993.

[3] K.P. van den Bos, H.C.L. Spelberg, A.J.M. Scheepstra, and J.R. de Vries, *De Klepel. Verantwoording, handleiding, diagnostiek en behandeling.*, Berkhout, Nijmegen, 1994.

[4] R.H.M. Brouwer, "Onderzoek naar de leesmoeilijkheid van Nederlandse proza," *Pedagogische Studiën*, vol. 40, no. 10, pp. 454–464, 1963.

[5] L. Ehri, *Handbook of reading research*, vol. II, chapter Development of the ability to read words, pp. 383–417, Longman, New York, 1991.

[6] D.J. Leu, "Oral reading error analysis: a critical review of research and application," *Reading Research Quarterly*, vol. 3, pp. 420–437, 1982.

[7] C.A. Chinn, M.A. Waggoner, R.C. Anderson, M. Schommer, and I.A.G. Wilkinson, "Situated actions during reading lessons - a microanalysis of oral reading error episodes," *American Educational Research Journal*, vol. 30, no. 2, pp. 361–392, 1993.

[8] J. Mostow, J. Beck, S.V. Winter, S. Wang, and B. Tobin, "Predicting oral reading miscues," in *Proc. of the 7th International Conference on Spoken Language Processing*, Denver, USA, 2002, pp. 1221–1224.

[9] J. Duchateau, M. Wigham, K. Demuynck, and H. Van hamme, "A flexible recogniser architecture in a reading tutor for children," in *Proc. of the ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, May 2006, pp. 59–64.

[10] Y.-C. Tam, J. Mostow, J. Beck, and S. Banerjee, "Training a confidence measure for a reading tutor that listens," in *Proc. European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 2003, pp. 3161–3164.

[11] K. Lee, A. Hagen, N. Romanyshyn, S. Martin, and B. Pellom, "Analysis and detection of reading miscues for interactive literacy tutors," in *Proc. International Conference on Computational Linguistics, Coling*, Geneva, Switzerland, August 2004.