



# Automatic Assessment of Children's Reading Level

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

Katholieke Universiteit Leuven, Belgium

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

## Abstract

In this paper, an automatic system for the assessment of reading in children is described and evaluated. The assessment is based on a reading test with 40 words, presented one by one to the child by means of a computerized reading tutor. The score that expresses the child's reading performance is calculated as the total time needed to read the 40 words divided by the number of correctly read words. In each grade, children are classified in 5 groups based on their score as provided by human annotators. We show that when the score for a child is assessed automatically using a speech recognizer, a classification can be obtained with a substantial agreement (Cohen's Kappa over 0.6) with the human classification. As all children in the experiments were classified either correctly or in an adjoining group, we can conclude that the proposed system can provide large time gains in current manual classification procedures.

**Index Terms:** computer aided language learning, reading assessment, ASR for children.

## 1. Introduction

One of the aims of the SPACE project<sup>1</sup> is the development of an automated reading tutor. The current version of the tutor contains tools for the management of reading tests, students and recordings, tools for the assessment of children, and remedial tools that provide feedback to the child.

From the point of view of the automatic speech recognition system used in the tutor, the main tasks to support these tools are detection and classification of reading errors, and tracking of the child's progress while reading. Mainly two facts make these tasks particularly hard: the reading capabilities of the envisaged young children are still rather weak, and for some tools like the tracking, the recognizer should work with very low latency (and of course real time). This paper focuses on the assessment tool in the current tutor, which provides an overall reading level assessment for a child.

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. Although

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

the quality of these instruments isn't questioned, it is clear that the administration of these tests is very time consuming. So, one of the aims of the proposed automatic reading level assessment is to provide a tool that allows a time gain in the current manual procedures.

This paper is organized as follows: in section 2, the Chorec database, used for the experiments, is described. The manual reading level assessment procedure is explained in section 3. Section 4 details the recognition system and the automatic reading error detection. The automatic reading level assessment is described and evaluated in section 5. Finally, in section 6, some conclusions and ideas for future work are given.

## 2. The Chorec database

The Chorec database contains reading sessions of Dutch-speaking elementary school children from grade 1 to 4. Until now about 300 children were recorded in 5 different schools, one of which is a school for children with known reading disabilities. For the experiments in this paper, the recordings from the 80 2nd grade children were used.

For every child, a computerized reading test battery is administered, containing real word reading tests, pseudoword reading tests and story reading tests. The reading level assessment described in this paper is based on the real word and pseudoword reading tests. There are 3 real word and 3 pseudoword reading tests with respectively 40 1-syllable, 40 2-syllable and 40 3- or 4-syllable real words or pseudowords. For some children, mainly from the school for children with known reading disabilities, the tests with 3- or 4-syllable words are not recorded as they are too difficult for the child.

All recordings have been transcribed and annotated manually, providing orthographic and phonetic transcriptions, and reading strategy and reading error annotations. From these transcriptions, the number of correctly read words within a test (i.e. the number of words correctly read in a child's last try), needed for scoring the child, was deduced. More details about the Chorec database and its annotations can be found in [1].

## 3. Manual reading level assessment

As is the case in most - if not all - reading tests, the real word reading test and the pseudoword reading test being used here are partially speed and partially accuracy tests. On the one hand, these are speed tests without a time limit, allowing the children to complete the test. On the other hand, accuracy matters too and must therefore also be taken into account [2].

As in many types of perceptual-motor tasks, there is a speed-accuracy trade-off, which is a trade-off between how fast a task can be performed and how many mistakes are made in performing the task. That is, a child can either perform the task very fast with a large number of errors or very slowly with very

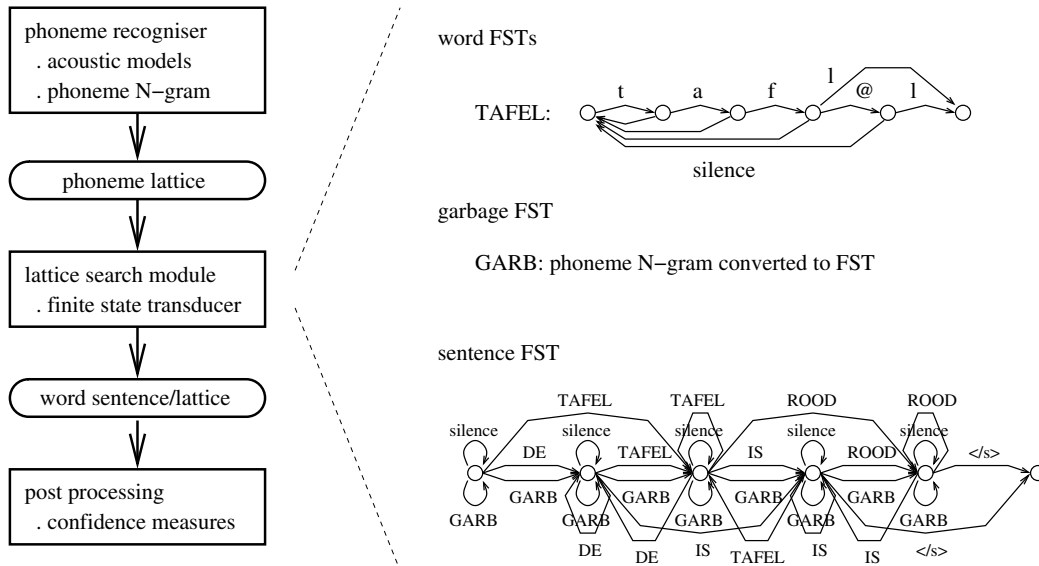


Figure 1: Recognition system architecture with 2 layers

few errors. When asked to perform a task as well as possible, children can apply strategies that may optimize speed, optimize accuracy, or combine both. For this reason, comparing the performance of two children cannot be done on the basis of speed or accuracy alone, but both values need to be taken into account [3]. The score we use is calculated as a mean-time-per-correct-item score; this comprises that the total response time is divided by the number of correctly read words.

Based on the distribution of this mean-time-per-correct-item score, percentiles or standard scores are used to discriminate between different performance levels. In the Netherlands and Flanders, CITO (Central Institute for Test Development) introduced the use of 5 performance groups: A: best performing 25%, B: above average performing 25%, C: below average performing 25%, D: far below average performing 15%, and E: worst performing 10%. These performance groups are directly and linearly derived from the percentile scores.

#### 4. Automatic reading error detection

In order to calculate a child's score, two values are needed: the total time used to read the 40 words in the test, and the number of errors made (or the number of correctly read words).

For the Chorec recordings, made by means of an automated reading tutor, the total time is easily found as the time is logged on which the screens appear and disappear.

To assess the number of errors made, an automatic reading error detection system was developed, based on a speech recognizer. This section describes and evaluates this system.

##### 4.1. Speech recognition system

The acoustic modeling is based on a 22 hour read speech database in Dutch, different from the Chorec database. It contains 16 kHz recordings of continuous sentences read by children aged between 5 and 11 years. Cross word context dependent acoustic models were estimated with 1400 tied (HMM) states and 16000 tied gaussian distributions in total. A straight-forward Mel-spectrum based signal processing scheme was adhered to, including cepstral mean subtraction, discriminant analysis and Vocal Tract Length Normalization (VTLN). When processing the Chorec data, recorded at 22050 Hz, the spectrum

over 8 kHz is used for VTLN warping factors over 1.0.

For the decoding of the speech, a recognition system architecture with two layers was adopted, as depicted in figure 1. This architecture can also be used for other recognition tasks. It was, for instance, applied successfully in large vocabulary continuous speech recognition as shown in [4].

In the first layer, a task independent phoneme recognizer generates a phoneme lattice. The phoneme recognizer is based on the acoustic modeling detailed above, and on a trigram phoneme sequence model estimated from a Dutch database with (correctly) read sentences.

The search engine in the second layer turns the phoneme lattice into a word level recognition result. This result may be a lattice that can be used in further processing. In the experiments in this paper, the recognition result is one sentence which allows to detect the reading errors. In the second layer, the task dependent information is modeled. As the sentence (or the word in the experiments below) that should be read is known in a reading tutor, we opted for a finite state transducer (FST) to produce a detailed model for the speech to be expected from the child. The FST used is a *composed* FST: the word FSTs (top right in figure 1) and the garbage FST (middle right) are inserted at the right places in the sentence FST (bottom right). The Dutch sentence in the example is *De tafel is rood* (The table is red).

More details on this recognition system architecture and on the acoustic modeling can be found in [5].

##### 4.2. Error detection in Chorec

For a 40-word test in Chorec, our recognition system is executed 40 times with a single word sentence FST. As explained before, we say that a word is read correctly when the child reads the word correctly at its last try, else a reading error occurred. So if a recognition result ends with the word itself, the word is tagged as read correctly, else, as a reading error.

In the above recognition system, we set the penalty for entering the garbage FST in such a way that the word FST will be recognized if possible (meaning that the phoneme string for the word is available in the phoneme lattice). This means that the recognition result can be influenced to output the word FST more often by adjusting the phoneme recognizer settings so that

a larger phoneme lattice is produced. The larger the phoneme lattice, the more words will be tagged as read correctly so the estimated number of correctly read words will increase. This means that the reading error detection rate (the percentage of reading errors correctly tagged as an error) will drop, but also that the false alarm rate (the percentage of correctly read words erroneously tagged as errors) will decrease.

For the experiments below, we tried two phoneme recognizer settings. The settings resulting in the largest lattices provided the best automatic reading level assessment, so those results are reported. The drawback of large lattices is a slow recognition, for example the total assessment procedure using large lattices is about two times slower than real time on a recent single processor computer. Note that the proposed assessment is offline, so processing slower than real time is not prohibitive.

For these large lattices, a reading error detection rate of 56% and a false alarm rate of 13% was found on the real words, for pseudowords these values are 74% and 26%. These results are worse than the ones presented in [5] for real words (in sentences), mainly for two reasons. First, in [5], the task was to detect miscues, which are defined as words that are not pronounced correctly *at once*. Miscues are easier to detect, for instance when the child needs several tries to read the word. Second, in [5], training and test material was extracted from the same database, this way avoiding a possible mismatch between both, and the recording conditions of the test data were better. For the Chorec database, recordings were made in normal rooms in schools, resulting in clearly audible background noise (nearby playground, traffic), speech by the adult assistant, a few recordings with a bad overall signal to noise ratio, ... We did not (manually) filter the Chorec test data for good recordings as we can expect the same data quality when the proposed automatic reading level assessment is used in practice.

## 5. Automatic reading level assessment

Based on a system with the error detection and false alarm rates presented in the previous section, it's difficult for an automated tutor to provide feedback to the child about each word separately. Given that 11% of the real words and 33% of the pseudowords are read incorrectly, a tutor that would point the child to every word that was detected as a reading error, would provide incorrect feedback about half of the time (more precisely for 64% and 42% of real words and pseudowords respectively).

However in this section we show that the described error detection system can provide useful information on reading errors when evaluated over a full 40-word test.

### 5.1. Assessment of the child's score

The better the number of correctly read words in a 40-word test is estimated, the better the estimated score for a child will be so the better its classification. The left part of figure 2 plots correct versus estimated number of correctly read words for the 80 2-syllable real word test recordings available in Chorec (see section 2). The right part relates the corresponding scores for the children: the human score, derived from the human annotations in Chorec, and the estimated score that is based on the estimated number of correctly read words. Figure 3 shows the same for the 2-syllable pseudoword test.

For the 2-syllable real word test, we can see that the number of correctly read words is close to 40 for most of the children. Classification of the children based on this value alone would be problematic. Fortunately, the use of both the total time needed

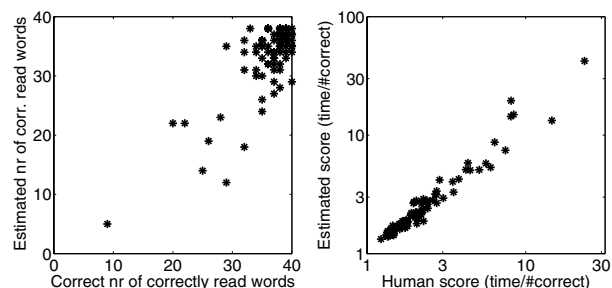


Figure 2: Correlation for 2-syllable real word test

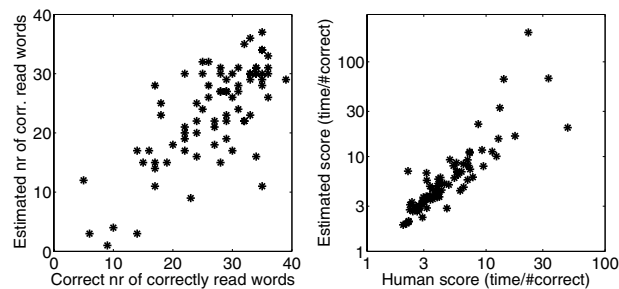


Figure 3: Correlation for 2-syllable pseudoword test

and the number of correctly read words in the score for the child seems to solve this problem.

The figure for the 2-syllable pseudoword test shows that this task is much more difficult for children. In this case, even the number of correctly read words alone can distinguish between children to a large extent. However, the correlation between the scores seems to be smaller, showing more outliers, so we can predict a worse classification based on the pseudoword test than on the real word test.

### 5.2. Assessment of the classification

For each of the 6 40-word tests, the children for which this test was recorded are divided into 5 performance groups based on their human scores. The groups range from A (best) to E (worst) as described in section 3. So for each test, this results in 4 limiting values for the scores between groups A and B, B and C, etc. Given these limiting values, the estimated score for a child can be converted easily to the corresponding performance group.

In table 1, an evaluation of the automatic reading level classification is given for each of the available word tests. The quality of the automatic classification is assessed in several ways: by the percentage of correctly classified children, by the percentage of children for which human and automatic classification differ by more than one performance group (a group A child classified as C, D or E etc.), and by Cohen's Kappa, both unweighted and with a linear weighting<sup>2</sup>. Cohen's Kappa with linear weighting reflects the distances between the ordinal classes we deal with: classifying A as C is worse than classifying A as B. For Kappa values between 0.6 and 0.8, the classifiers show a *substantial* agreement, Kappa values over 0.8 are said to reflect an *almost perfect* agreement. For all Kappa values in the table, the standard deviation is about 0.06.

From table 1 we can conclude that for the real word tests, there is a substantial agreement between human and automatic classification. Agreement based on the pseudoword tests is slightly worse, as we already expected from the analysis of the

<sup>2</sup>See e.g. <http://faculty.vassar.edu/lowry/kappaexp.html>

	correct	err. > 1	unw. $\kappa$	lin. $\kappa$
1-syll. real	73.4%	0.0%	0.66	0.82
2-syll. real	70.9%	0.0%	0.63	0.80
3+4-syll. real	65.8%	1.4%	0.57	0.77
1-syll. pseudo	54.4%	1.3%	0.42	0.67
2-syll. pseudo	62.8%	3.8%	0.53	0.71
3+4-syll. pseudo	42.0%	14.5%	0.28	0.49

Table 1: Evaluation of the baseline automatic classification

	correct	err. > 1	unw. $\kappa$	lin. $\kappa$
1-syll. real	75.9%	0.0%	0.69	0.84
2-syll. real	80.0%	0.0%	0.74	0.86
3+4-syll. real	68.5%	0.0%	0.60	0.79
1-syll. pseudo	70.9%	1.3%	0.63	0.78
2-syll. pseudo	67.9%	2.6%	0.59	0.76
3+4-syll. pseudo	51.7%	1.7%	0.30	0.50

Table 2: Evaluation of classification with bias compensation

underlying scores in the previous section. It should be noted that slightly worse results for both 3+4-syllable tests can be expected because for some children with weak reading capabilities, these tests were not recorded. We could include them by giving a very high score (corresponding to zero words correctly read). But also the automatic scoring would be easy for these children, resulting in a correct classification anyway. Hence we excluded them when assessing the automatic reading level classification. However this also implicates that the resulting range of reading levels is smaller for the 3+4-syllable tests so classification based on the 3+4-syllable tests is slightly harder.

In order to get a better idea about what good and bad Kappa values are for this classification task, we can compare the values in the table with Kappa values between human classifications based on two different real word tests or between human classifications based on two different pseudoword tests. It is expected, for instance, that a child that scores well on the 2-syllable real word test, will also score well on the 3+4-syllable real word test. Comparing human classifications on different real word tests, unweighted Kappa values lie between 0.34 and 0.56 (with linear weighting between 0.55 and 0.74). For pseudoword tests, unweighted Kappa values range from 0.40 to 0.44 (with linear weighting values lying between 0.61 and 0.68). From these values we can conclude that the obtained agreements for automatic classification are substantial indeed.

We observed that on average for each word test, the estimated number of correctly read words is lower than the correct number. As explained in section 4, this can be solved by using even larger phoneme lattices. However, recognition time then increases substantially so that this solution is not useful in practice. Therefore, we decided to compensate for this bias towards too low automatic estimates for the number of correctly read words by adding the (word test dependent) bias to the estimate. It should be noted that when correcting the estimated number of correctly read words for a child, the bias is to be calculated on the other children only (leaving-one-out parameter estimation).

The results with bias compensation are given in table 2, they are slightly better than the ones without bias compensation. When looking at the number of children with a difference between human and automatic classification larger than one, we see that using the pseudoword tests, this very bad classification occurs for 1 or 2 children. Using real word tests however, all children are classified either correctly or in a neighboring performance group.

Based on this finding, the efficiency can be enhanced of the current manual reading level classification of children. For instance, as for remedial purposes the attention goes out to children in performance group E, only children classified automatically in groups D and E, this is 25% of the total, need to be checked manually: a gain in time by a factor 4.

It should be noted that for making the Chorec recordings used in the experiments, the child was assisted by an adult who decided when to proceed to the next screen with the next word. In the current reading tutor, this decision can be made automatically based on the outcome of an automatic speech recognition system like the one used in this paper for reading error detection. However, we don't know how this automatic progression of the screens influences the reading level classification accuracy. Still, the current manual scoring procedure can't be executed in real time while listening to the child reading, so the time gain remains.

## 6. Conclusions and future work

In this paper, we have shown that reading level classification of second grade children can be done automatically with substantial agreement with the human classification. As all children can be classified either correctly or in an adjoining performance group, we can also conclude that the proposed system can be used to provide large time gains in current manual reading level classification procedures.

Future work lies in the improvement of the automatic reading error detection system so that other uses may become possible, like for instance direct feedback to the child in an automated reading tutor. One way to improve the reading error detection is by post processing the current recognition results, using new information sources like the word identity, prosodic cues, or segment duration.

## 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] Leen Cleuren, Jacques Duchateau, Alain Sips, Pol Ghesquière, and Hugo Van hamme, "Developing an automatic assessment tool for children's oral reading," in *Proc. IC-SLP*, Pittsburgh, U.S.A., Sept. 2006, pp. 817–820.
- [2] Margo G.H. Jansen, "The Rasch model for speed tests and some extensions with applications to incomplete designs," *Journal of Educational and Behavioral Statistics*, vol. 22, no. 2, pp. 125–140, 1997.
- [3] Ian Dennis and Jonathan St.B.T. Evans, "The speed-error trade-off problem in psychometric testing," *British Journal of Psychology*, vol. 87, pp. 105–129, 1996.
- [4] Kris Demuyneck, Dirk Van Compernelle, and Hugo Van hamme, "Robust phone lattice decoding," in *Proc. IC-SLP*, Pittsburgh, U.S.A., Sept. 2006, pp. 1622–1625.
- [5] Jacques Duchateau, Mari Wigham, Kris Demuyneck, and Hugo Van hamme, "A flexible recogniser architecture in a reading tutor for children," in *Proc. ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, May 2006, pp. 59–64.