Automatic Classification of Reading Disorders in a Single Word Reading Test

Andreas Maier
Abteilung für Phoniatrie und Pädaudiologie,
Universitätsklinikum Erlangen
Bohlenplatz 21
Erlangen, Germany
andreas.maier@cs.fau.de

Stefanie Horndasch
Kinder- und Jugendabteilung für Psychische Gesundheit in der Psychiatrischen und Psychotherapeutischen Klinik Erlangen
Schwabachanlage 6
Erlangen, Germany
stefanie.horndasch@uk-erlangen.de

Elmar Nöth
Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
Martensstr. 3
Erlangen, Germany
elmar.noeth@cs.fau.de

ABSTRACT

In clinical practice, reading disorders are still evaluated perceptually. In order to alleviate this problem, we propose to use automatic speech processing techniques to classify reading disorders. Therefore, we recorded 38 children who were suspected to have a reading disorder. The recordings were performed using a German standard test for reading disorders. Each child was perceptually assessed and the number of reading errors per child was recorded. Furthermore, the reading duration was stored for each child. If either of both values exceeded an age-dependent limit, the child was diagnosed having a reading disorder. In 30 of the 38 children the reading disorder was confirmed. In this paper, we present the results on the automatic evaluation concerning a single word reading test. We achieve up to 78.9 % recognition rate in the detection of the exceedance of the reading error limit and 97.4 % recognition rate in the classification for reading disorder.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement, Performance

Keywords

Reading Disorders, Automatic Speech Processing, Automatic Reading Assessment

1. INTRODUCTION

The state-of-the-art approach to examine children for reading disorders is a perceptual evaluation of the children's reading abilities. In all of these reading tests, a list of words or sentences is presented to the child. The child has to read all of the material as fast and as accurately as possible. In order to determine whether the child has a reading disorder two variables are investigated by a human supervisor during the test procedure:

- The duration of the test, i.e. the fluency, and
- The number of reading errors during the reading of the test material, i.e., the accuracy.

Both variables, however, are dependent on the age of the child and related to each other. If a child tries to read very fast, the number of reading errors will increase and vice versa [2]. Furthermore, with increasing age the reading ability of children increases. Hence, appropriate test material has to be chosen according to the age and reading ability of the child. Therefore, reading tests often consist of different sub-tests. While younger children are tested with really existing words and only short sentences, the older children have to be tested with more difficult tasks, such as long complex sentences and pseudo words which may or may not resemble real words. Appropriate sub-tests are then selected for each tested child. Often this is linked to the child's progress in school.

One major drawback of the testing procedure is the intra-rater variablity in the perceptual evaluation procedure. Although the test manual often defines how to differentiate reading errors from normal disfluencies and "allowed" pronunciation alternatives, there is no exact definition of a reading error in terms of its acoustical representation. In order to solve this problem, we propose the use of a speech recognition system to detect the reading errors. This procedure has two major advantages:

- The intra-rater variability of the speech recognizer is zero because it will always produce the same result given the same input.

- The definition of reading errors is standardized by the parameters of the speech recognition system, i.e., the reading ability test can also be performed by lay persons with only little experience in the judgment of readings disorders.

In the literature, different automatic approaches to determine the “reading level” of a child exist. Often the reading level is linked to the perceptual evaluation of expert listeners using five to seven classes. In [1] Black et al. estimate a reading level between 1 and 7 using pronunciation verification methods based on Bayesian Networks. Compared to the human evaluation they achieve correlations between their automatic predictions and the human experts of up to 0.91 on 13 speakers. In [3] the use of finite-state-transducers is proposed to obtain a “reading level” between “A” (best) and “E” (worst). For this five-class problem absolute recognition rates of up to 73.4 % for real words and 62.8 % for pseudo words are reported. In order to remove age-dependent effects from the data, 80 children in the 2nd grade were investigated. Both papers focus on the creation of a “reading tutor” in order to improve children’s reading abilities.

In contrast to these studies, we are interested in the diagnosis of reading disorders as they are relevant in a clinical point of view. Currently, we are developing PEAKS (**P**rogram for the **E**valuation of **A**ll **K**inds of **S**peech Disorders [10]) — a client-server-based speech evaluation framework — which was already used to evaluate speech intelligibility in children with cleft lip and palate [11], patients after removal of laryngeal cancer [7], and patients after the removal of oral cancer [14]. PEAKS features interfaces and tools to integrate standardized speech tests easily. After integration of a new test, PEAKS can be used for recording from any PC which is connected to the Internet if Java Runtime Environment version 1.6 or higher is installed. All analyses performed by PEAKS are fully automatic and independent of the supervising person. Hence, it is an ideal framework to integrate an automatic reading disorder classification system.

In the following we describe the used recognition system, followed by a description of the speech data. In the results section we present the outcome of the automatic classification procedure. The paper is concluded by a summary.

2. METHODS

The “Program for the Evaluation of All Kinds of Speech disorders” (PEAKS, [11]) was applied to record and analyze the data. It is an online speech recording and evaluation system which is available via the Internet (http://peaks.informatik.uni-erlangen.de).

In a first acoustic analysis, the speech recognizer converts speech into a sequence of feature vectors which consist of 12 Mel-Frequency Cepstrum Coefficients (MFCC). The first coefficient is replaced with the energy of the signal. Additionally 12 delta coefficients are computed over a context of 2 time frames to the left and the right side (50 ms in total).

Table 2: 38 Children were recorded with the SLRT: The table shows mean value, standard deviation, minimum, and maximum of the age of the children and the count (#) in the respective group.

group	#	mean	std. dev.	min	max
all	38	9.7	0.9	7.8	11.3
girls	12	10.2	0.7	9.0	11.3
boys	26	9.5	0.9	7.8	11.3

The recognition is performed with semi-continuous Hidden Markov Models. The codebook contains 500 full covariance Gaussian densities which are shared by all HMM states. The elementary recognition units are polyphones [12], a generalization of triphones. Polyphones use phones in a context as large as possible which can still statistically be modeled well, i.e., the context appears more often than 50 times in the training data. The HMMs for the polyphones have three to four states per phone.

From the recognized word chain and the reference a percentage of correctly uttered words — the word accuracy (WA) — is computed. As previously shown, this number corresponds with the speech intelligibility [13], i.e., since the recognizer and the setup is kept constant the only varying factor remains the child’s speech that is recorded [9].

Furthermore, the system keeps track of the recording time, i.e, the reading duration. This number is also used as a feature in the following.

3. DATA

3.1 Test Material

The recorded test data is based on a German standardized reading disorder test — the “Salzburger Lese-Rechtschreib-Test” (SLRT, [8]). In total the SLRT consists of eight subtests (cf. Table 1). All sub-tests contain 196 words of which 170 are disjoint.

The test is standardized according to the instructions and the evaluation. The test is presented in form of a small book, which is handed to the children to read in. They get the instruction to read the text as fast as possible while doing as little reading mistakes as possible. In the original setup, the supervisor of the test has to measure the time for all sub-tests separately while noting down the reading errors of the child.

In the following, we will only report the results obtained for the SLRT2 sub-test. It is composed of a list of 30 mono- and bisyllabic real words, i.e., a single word test.

3.2 Recording Setup

In order to be able to collect the data directly at the PC, the test had to be modified. Instead of a book, the text was presented as a slide on the screen of a PC. The instructions to the child were the same as in the original setup.

All children were recorded with a head-mounted micro-

Table 1: Structure of the SLRT test: The table reports all sub-tests of the SLRT with their contents, their number of words, and the school grades in which the respective sub-test is suitable.

sub-test	content	# of words	grade
SLRT1	A short list of bisyllabic, single, real words to introduce the test. This part is not analyzed according to the protocol of the test.	8	1–4
SLRT2	A list of mono- and bisyllabic real words	30	1–4
SLRT3	A list of compound words with two to three compounds each	11	3–4
SLRT4	A short story with only mono- and bisyllabic words	30	1–2
SLRT5	A longer story with mainly mono- and bisyllabic words but also a few compound words	57	3–4
SLRT6	A short list of pseudo words with two to three syllables to introduce the pseudo words. This part is not analyzed according to the protocol of the test.	6	3–4
SLRT7	A list of pseudo words with two to three syllables	24	1–4
SLRT8	A list of mono- and bisyllabic pseudo words which resemble real words	30	2–4

Table 3: Overview on the limits of pathology for the SLRT7 and SLRT8 sub-tests

	SLRT 2	
grade	# of errors	duration [s]
1st	6	119
2nd	4	76
3rd	2	33
4th	2	29

phone (Plantronics USB 510) at the University Clinic Erlangen. The recordings took place in a separate quiet room without background noises. Hence, appropriate audio quality was achieved in all recordings.

In total 38 children (26 boys and 12 girls) were recorded. The average age of the children was 10.2 ± 0.9 years. A detailed overview regarding the statistics of the children's ages is given in Tabelle 2. All of the children were speculated to have a reading disorder.

3.3 Perceptual Evaluation

For each child the decision whether its reading ability was pathologic or not was determined according to the manual of the SLRT [8]. A child's reading ability is deemed pathologic

- if the duration of the test is longer than an age-dependent standard value or
- if the number of reading errors exceeds an age-dependent standard value.

These limits differ for each sub-test according to the SLRT. Table 3 reports these limits for the sub-test SLRT2. In the SLRT2 sub-test 30 children were above the time limit.

We assigned each child two different labels: "reading error/normal" and "pathologic/non-pathologic". If only the number of misread words is exceeded, the child is assigned the label "reading error", otherwise "normal". Reading errors are regarded as soon as a single phonemic deviation is found. Errors of the accentuation of the word are also counted as reading errors as described in the manual of the test [8]. In total 9 children exceeded the error limit.

If either of these two boundaries is exceeded by the child, the child is assigned the label "pathologic". 30 of the 38 children were diagnosed to have pathologic reading.

4. RESULTS

Classification was performed in a leave-one-speaker-out (LOO) manner since there was only little training and test data available. We chose two popular measures in order to report the classification accuracy: The recognition rate (RR) and the area under the Receiver-Operating-Characteristic (ROC) curve [4]. As classification system, we decided for Ada-Boost [5] in combination with an LDA-Classifier as weak learner as it was already successfully applied in [6].

In total we employed four different features for the recognition task:

- word accuracy (WA) of the children's speech recognition system
- duration of the test measured in milliseconds
- the reading error limit and the duration limit for the respective age of the child
- the age of the child at the time of the recording in months

Table 4: Results for the both classification tasks “reading error/normal” and “pathologic/non-pathologic”.

task:	“reading error”		“pathologic”	
feature	ROC	RR [%]	ROC	RR [%]
WA	0.536	68.4	0.654	68.4
+ duration	0.609	71.1	0.996	94.7
+ age-dependent limits	0.552	68.4	0.996	94.7
+ age	**0.625**	**78.9**	**0.999**	**97.4**

In order to explore the effect of the different features we started our experiments using WA only and added more features. Table 4 reports the results. The best classification performance is found for both classification tasks as a combination of recognition accuracy, age-dependent limits, age of the child, and the reading duration. The performance for the task “reading error” with an area under the curve of 0.625 is moderate. Also the classification rate of 78.9 % shows moderate to good performance. The detection of the reading pathology, however, works very well with the automatic system with an area under the curve of 0.999 and a classification rate of 97.4 %. With such an performance our system almost matches the human evaluator exactly. Note, that the presented system does not require any human post-processing at all. If similar results are achieved in the other sub-tests a fully automatic detection of reading disorders will be possible.

5. SUMMARY

We presented a novel approach for the classification of reading disorders using automatic speech recognition techniques. The detection rate of reading errors was moderate with 78.9 %. The classification of the pathology could be achieved with 97.4 %. In the future, this technique may facilitate the work of speech and language therapists.

6. REFERENCES

[1] M. Black, J. Tepperman, S. Lee, and S. Narayanan. Estimation of children’s reading ability by fusion of automatic pronunciation verification and fluency detection. In *Interspeech 2008 – Proc. Int. Conf. on Spoken Language Processing, 11th International Conference on Spoken Language Processing, September 25-28, 2008, Brisbane, Australia, Proceedings*, pages 2779–2782, 2008.

[2] I. Dennis and J. S. B. T. Evans. The speed-error trade-off problem in psychometric testing. *British Journal of Psychology*, 87:105–129, 1996.

[3] J. Duchateau, L. Cleuren, H. V. Hamme, and P. Ghesquiere. Automatic assessment of children’s reading level. In *Interspeech 2007 – Proc. Int. Conf. on Spoken Language Processing, 10th European Conference on Spoken Language Processing, August 27-31, 2007, Antwerp, Belgium, Proceedings*, pages 1210–1213, 2007.

[4] A. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[5] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156, San Mateo, CA, USA, 1996. Morgan Kaufmann.

[6] C. Hacker, T. Cincarek, A. Maier, A. Heßler, and E. Nöth. Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 197–200, Hawaii, USA, 2007. IEEE Computer Society Press.

[7] T. Haderlein, K. Riedhammer, A. Maier, E. Nöth, H. Toy, and F. Rosanowski. An Automatic Version of the Post-Laryngectomy Telephone Test. In V. Matoušek and P. Mautner, editors, *10th International Conf. on Text, Speech and Dialogue (TSD)*, volume 4629 of *Lecture Notes in Artificial Intelligence*, pages 238–245, Berlin, Heidelberg, New York, 2007. Springer.

[8] K. Landerl, H. Wimmer, and E. Moser. *Salzburger Lese- und Rechtschreibtest. Verfahren zur Differentialdiagnose von Störungen des Lesens und des Schreibens für die 1. bis 4. Schulstufe*. Huber, Bern, 1997.

[9] A. Maier. *Speech of Children with Cleft Lip and Palate: Automatic Assessment*. Logos Verlag, Berlin, Germany, 2009.

[10] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth. PEAKS – A System for the Automatic Evaluation of Voice and Speech Disorders. *Speech Communication*, 51(5):425–437, 2009.

[11] A. Maier, E. Nöth, A. Batliner, E. Nkenke, and M. Schuster. Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate. *Informatica*, 30(4):477–482, 2006.

[12] E. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Automatic Speech Recognition without Phonemes. In *Proceedings European Conference on Speech Communication and Technology (Eurospeech)*, pages 129–132, Berlin, Germany, 1993.

[13] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth. Evaluation of Speech Intelligibility for Children with Cleft Lip and Palate by Automatic Speech Recognition. *Int J Pediatr Otorhinolaryngol*, 70:1741–1747, 2006.

[14] M. Windrich, A. Maier, R. Kohler, E.Nöth, E. Nkenke, U. Eysholdt, and M. Schuster. Automatic Quantification of Speech Intelligibility of Adults with Oral Squamous Cell Carcinoma. *Folia Phoniatr Logop*, 60:151–156, 2008.