An Automatic Version of a Reading Disorder Test

ANDREAS MAIER, FLONAN HÖNIG, STEFAN STEIDL, and ELMAR NÖTH,
Universität Erlangen-Nürnberg
STEFANIE HORNDASCH, ELISABETH SAUERHÖFER, OLIVER KRATZ,
and GUNTHER MOLL, Universitätsklinikum Erlangen

We present a novel system to automatically diagnose reading disorders. The system is based on a speech recognition engine with a module for prosodic analysis. The reading disorder test is based on eight different subtests. In each of the subtests, the system achieves a recognition accuracy of at least 95%. As in the perceptual version of the test, the results of the subtests are then joined into a final test result to determine whether the child has a reading disorder. In the final classification stage, the system identifies 98.3% of the 120 children correctly. In the future, our system will facilitate the clinical evaluation of reading disorders.

Categories and Subject Descriptors: H.4.M [**Information Systems Applications**]: Miscellaneous

General Terms: Measurement, Performance

Additional Key Words and Phrases: Reading disorders, automatic speech processing, automatic reading assessment

ACM Reference Format:
Maier, A., Hönig, F., Steidl, S., Nöth, E., Horndasch, S., Sauerhöfer, E., Kratz, O., and Moll, G. 2011. An automatic version of a reading disorder test. ACM Trans. Speech Lang. Process. 7, 4, Article 17 (August 2011), 15 pages.
DOI = 10.1145/1998384.1998391 http://doi.acm.org/10.1145/1998384.1998391

1. INTRODUCTION

In the last few years the use of automatic speech recognition (ASR) techniques for clinical purposes became more and more popular [Kitzing et al. 2009]. Before the application of ASR to clinical use, the therapists and clinicians evaluated their patients' speech using perceptual means only. As any perceptual evaluation, this process suffers from inter- and intraindividual differences. For scientific studies, the perceptual evaluation is either performed by a panel of experts where the mean opinion is chosen as a reference, or standardized values are obtained using a standardized test protocol and many hundred or thousand control subjects. With ASR many of the time- and manpower-consuming perceptual tests could be replaced by automatic means. This was achieved for:

—intelligibility of read speech in patients with oral cancer [Maier et al. 2007] and patients after removal of the larynx [Haderlein 2007] in German;
—speech of children with cleft lip and palate in a pictogram naming task concerning their intelligibility [Maier et al. 2009a] and their specific speech disorders [Maier et al. 2008] in German and Italian language [Scipioni et al. 2009];

Authors' address: Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg Martensstrasse 3, 91058, Germany; email: Andreas.maier@cs.fau.de.

DOI 10.1145/1998384.1998391 http://doi.acm.org/10.1145/1998384.1998391

—general assessment of speech intelligibility using a randomized syllable reading task in Dutch [Middag et al. 2009];
—the visualization of speech [Maier et al. 2009d] and voice disorders [Haderlein et al. 2006] for clinical purposes [Maier et al. 2009e].

Hence, ASR is a powerful tool that can serve clinical purposes well.

1.1. Clinical Evaluation of Reading Disorders

Reading disorders are complex and have been researched extensively in the past in the field of psychiatry [Beitchman and Young 1997]. 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, that is, the fluency, and
—the number of reading errors during the reading of the test material, that is, the accuracy.

Both variables, however, are dependent on the age of the child and are related to each other. If a child tries to read very fast, the number of reading errors will increase, and conversely, when reading slowly, fewer errors will occur [Dennis and Evans 1996]. 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 subtests. While younger children are tested with meaningful words and only short sentences, the older children have to be tested with more difficult tasks, such as long complex sentences and pseudowords that may or may not resemble real words. Appropriate subtests 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- and interrater variability 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 resolve this issue, the human observers have to be trained before they are able to perform the testing. In the test material that we chose as reference, the test setup was standardized with several thousand children. Each new test result is compared to a list of standard values that enables the comparison of the tested child to the standard set. In this manner, a percentage rank is obtained. Based on these extensive statistics, the decision whether a child's reading ability is disordered or normal is then made. This procedure is considered as the state of the art of clinical evaluation of reading disorders.

Intra- and interrater variability is removed, if the test is not based on perceptual evaluation. Hence, we propose the use of a speech recognition system to detect reading errors. This procedure has two major advantages.

—The intrarater 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, that is, the reading ability test can also be performed by lay persons with only little experience in the judgment of reading disorders. The testing can be performed by a lay person. No expensive training of a human observer is required.

1.2. State of the Art in Reading Level Assessment and Reading Tutoring

Many automatic approaches on reading assessment and tutoring exist. In this section we provide a few of the many successful examples that are found in the literature to introduce the reader to this topic.

A reading tutor is a system that processes read speech in real-time. It is designed to aid children in the training of their reading ability by provision of meaningful feedback. Aist [2000] showed that such an automatic system can be very effective for certain tasks, like the training of word comprehension. For this task, there was no significant difference between the automatic system and a human reading tutor.

More recent versions of reading tutors try to elicit speech of children using a dialogue strategy. One of the most challenging points in the design of such a system is that the dialogue should be technically feasible on the one hand and educationally effective on the other hand. Hence, a trade-off has to be found that restricts the responses of the children to a certain task domain. An example for this can be found in Aist and Mostow [2009], where an attempt is made to teach the human learner an implicit grammar that is compatible with the automatic system.

To determine the "reading level" of a child usually a short text passage has to be read by the child. Often the reading level is linked to the perceptual evaluation of expert listeners using five to seven classes. Black et al. [2008] estimate a reading level between 1 and 7 using pronunciation verification methods based on Bayesian Networks. They achieve correlations between their automatic predictions and the human experts of up to 0.91 on 13 speakers. In Duchateau et al. [2007] 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 pseudowords are reported. In order to remove age-dependent effects from the data, 80 children in 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 [Maier et al. 2009a]) a client-server-based speech evaluation framework that was already used to evaluate speech intelligibility in children with cleft lip and palate [Maier et al. 2006b], patients after removal of laryngeal cancer [Schuster et al. 2006], and patients after the removal of oral cancer [Windrich et al. 2008]. 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 that is connected to the Internet if Java Runtime Environment version 1.6 or higher is installed. All analyzes 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.

The article is organized as follows. First the test material, the recorded speech data, and its annotation is described and discussed. Next, the automatic evaluation methods, that is, the speech recognizer and the classifiers, are introduced. In the results section the classification accuracy is presented in detail. The subsequent section discusses the outcome of the experiments. The paper is concluded by a summary.

2. SPEECH DATA

In order to be able to interpret the results and to compare them to other studies, the test material, speech data, and its annotation are described in detail here. Special attention is given to the annotation procedure since the automatic evaluation algorithm is aimed at being used for clinical diagnosis. Therefore, the annotation should meet clinical standards.

Table I.
Structure of the SLRT test: The table reports all subtests of the SLRT with their contents, their number of words, and the school grades for which the respective sub-test is suitable. Note that the school grade is highly correlated with the age of each child

sub-test	content	# of words	grade
SLRT1	A short list of bisyllabic, single, real words to introduce the test. This part is just used for introducing the material, not for diagnosis.	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 pseudowords with two to three syllables to introduce the pseudowords. This part is just used for introducing the material, not for diagnosis.	6	3–4
SLRT7	A list of pseudowords with two to three syllables	24	1–4
SLRT8	A list of mono- and bisyllabic pseudowords that resemble real words	30	2–4

2.1. Test Material

The recorded test data is based on a German standardized reading disorder test: the Salzburger Lese-Rechtschreib-Test (SLRT), as presented in Landerl et al. [1997]. In total the SLRT consists of eight subtests (cf. Table I). All subtests contain 196 words, of which 170 are unique.

The test is standardized according to the instructions and the evaluation procedure. The test is presented in the form of a small book that is handed to the children to read in. They are instructed to read the text as fast as possible while making as few reading mistakes as possible.

In the original setup the supervisor of the test has to measure the duration for all subtests separately while noting down the reading errors of the child. As the SLRT1 and SLRT6 subtests are only meant to introduce the type of test material to the children, both subtests are not evaluated.

2.2. 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 [Maier et al. 2009c]. The instructions to the child were the same as in the original setup.

All children were recorded with a head-mounted microphone (Plantronics USB 510). The microphone was placed approximately 2 to 3 cm away from the mouth and was covered with foam in order to prevent noises from breathing and loud plosives. The recordings took place in a separate, quiet room without background noises. The walls were mostly covered with furniture that reduced the reverberation time in the room. Hence, appropriate audio quality was achieved in all recordings. The control group of children was recorded at a local elementary school.

In total 120 children were recorded. All children were native German speakers, had normal hearing, and were using the same local dialect. Eighty-two of them were recorded at a local elementary school. In order to increase the number of children with reading pathology in our data set, we additionally recorded 38 children with a diagnosed reading disorder. The average age of the children was 9.6 ± 0.9 years. A detailed overview regarding the statistics of the children's ages is given in Table II. The relation between age and pathologic reading is reported in the results section.

Table II.
The SLRT test was recorded for 120 children: Due to age restrictions of the SLRT in subtest SLRT3 and SLRT8 fewer children were collected. 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
SLRT2, SLRT4+5, SLRT8	120	9.6	0.9	7.4	11.3
SLRT3	102	9.7	0.6	8.3	11.3
SLRT8	115	9.6	0.9	7.4	11.3

2.3. Perceptual Evaluation

In the following, we describe the human, that is, perceptual evaluation of the speech data. For each child, the decision whether his or her reading ability was pathologic or not was determined according to the values given in the manual of the SLRT [Landerl et al. 1997]. Although the manual presents standardized values for deciding whether a child has pathologic reading or not, the decision is still difficult.

—If the duration of the test is longer than the age-dependent error limit the child's reading ability is classified as "pathologic."
—If the duration is within the time limits, the number of reading errors is investigated. Reading errors are marked as soon as a single phonemic deviation is found. Frequent substitutions which are caused by dialectal influences and self-corrections were not counted as errors as described in the manual of the test [Landerl et al. 1997]. The manual offers an age-dependent standard value but the decision for or against a pathologic finding has to be made by the clinician. As this process is very difficult to model with an automatic system, we decided in this work to follow the SLRT manual and assign the label "pathologic" in all cases that exceeded the reading error limit. For all the 38 children with a prior diagnosed reading disorder, this process indeed results in a pathologic finding.

The time and error limits differ for each subtest according to the SLRT. For this work we defined the 10th percentile of the SLRT as the limit for pathology for the duration and the number of reading errors.

Table III reports the results of the different subtests. Note that the subtests SLRT4 and SLRT5 were joined into one group as their age limits are disjoint. No single subtest is able to identify all of the 38 children. The fusion of their results, however, is very effective: All of the children who were diagnosed with a reading disorder by our clinical partner were identified by at least one subtests of the SLRT as pathologic. No child of the 82 control children was identified as pathologic in any of the subtests.

3. AUTOMATIC EVALUATION SYSTEM

The automatic evaluation is based on four information sources:

—the total duration of the test,
—the reading error and duration limits,
—the word accuracy computed by a speech recognition system, and
—prosodic information

The test duration can be easily accessed as PEAKS tracks this information automatically during the recording. Prior information about the child, namely the child's age and the respective duration and error limits determined by the child's school grade, can also easily be obtained as it is known to the clinician.

Table III.
Overview on the result of the perceptual evaluation of the SLRT subtests. No subtest alone is able to identify all the 38 children with reading disorder. The combination, however, identifies all of the children as reading disordered

sub-test	# of "pathologic" children
SLRT2	30
SLRT3	26
SLRT4+5	34
SLRT7	31
SLRT8	30

3.1. Speech Recognition Engine

For the objective measurement of the reading accuracy, we use an automatic speech recognition system based on Hidden Markov Models (HMM). It is a word recognition system developed at the Pattern Recognition Lab (Lehrstuhl für Mustererkennung) of the University of Erlangen-Nuremberg. In this study, the latest version as described in detail in Gallwitz [2002] and Stemmer [2005] was used.

As features we use 11 Mel-Frequency Cepstrum Coefficients (MFCCs) and the energy of the signal plus their first-order derivatives. The short-time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10ms. The filter bank for the Mel-spectrum consists of 25 triangular filters. The 12 delta coefficients are computed over a context of 2 time frames to the left and the right side (56 ms in total).

The recognition is performed with semicontinuous HMMs. The codebook contains 500 full covariance Gaussian densities that are shared by all HMM states. The elementary recognition units are polyphones [Schukat-Talamazzini et al. 1993], a generalization of triphones. Polyphones use phones in a context as large as possible, with the restriction that that enough training data must be available for sound statistical modeling. Our heuristic is to use contexts that appear at least 50 times in the training data. The HMMs for the polyphones have three to four states.

We used a unigram language model to weigh the outcome of each word model. It was trained separately for each SLRT subtest on the respective reference text. For our purpose it was necessary to emphasize the acoustic features in the decoding process.

In Maier et al. [2009b] a comparison between different language models was conducted. It was shown that intelligibility can be predicted using word recognition accuracies computed from different language models. Higher-order language models were found to increase the recognition rate, at the cost of the relation between recognition result and speech intelligibility. We chose unigram language modeling, because it is computationally more efficient than a zerogram on the one hand and exhibits a sufficiently restricted linguistic contribution on the other. The size of the recognizer's lexicon was determined by the number of distinct words in the respective subtest.

The result of the recognition is a word sequence. In order to get an estimate of the quality of the recognition, the word accuracy (WA) is computed. Based on the number of correctly recognized words C and the number of words R in the reference, the WA is further dependent on the number of wrongly inserted words I:

$$\mathrm{WA} = \frac{C - I}{R} \cdot 100\%.$$

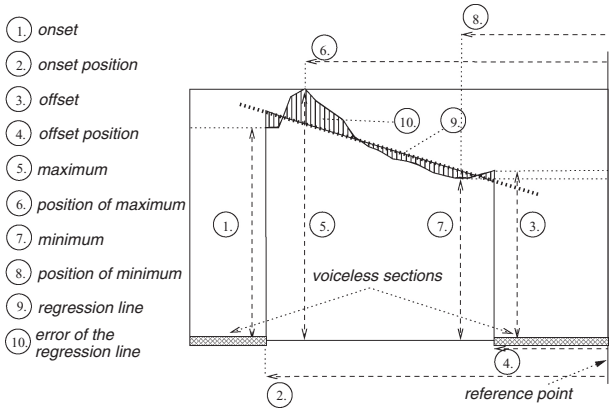

Fig. 1. Computation of prosodic features for the fundamental frequency contour within one word after Kiessling et al.

Hence, the WA can take values between minus infinity and 100%. Furthermore, we computed the word recognition rate (WR)

$$\mathrm{WR} = \frac{C}{R} \cdot 100\%$$

as it disregards insertions as an error. This measure does not punish self-repetitions and is known to be highly correlated to speech intelligibility [Maier et al. 2009a].

The speech recognition system had been trained on separate audio data from 23 male and 30 female children from a local school who were between 10 and 14 years old (6.9 hours of speech). To make the recognizer more robust, we added 2.3 hours of data from 85 male and 47 female adult speakers from all over Germany from the VERBMOBIL project [Wahlster 2000]. The data were recorded with a close-talk microphone with 16kHz sampling frequency and 16 bit resolution. The adult speakers were from all over Germany and thus covered most dialect regions. However, they were asked to speak standard German. The adults' data were adapted by vocal tract length normalization as proposed in Stemmer et al. [2003]. During training, a validation set was used that only contained children's speech. MLLR adaptation [Gales et al. 1996; Maier et al. 2006a] with the patients' test data led to further improvement of the speech recognition system.

3.2. Prosodic Features

The prosody module used in these experiments was originally developed within the VERBMOBIL project [Wahlster 2000], mainly to speed up the linguistic analysis [Nöth et al. 2000; Batliner et al. 2000; Kiessling 1997]. It assigns a vector of prosodic features to each word in a word sequence that is then used to classify a word with respect to, for instance, carrying the phrasal accent and being the last word in a phrase. For this paper, the prosody module takes the text reference and the audio signal as input and returns 37 prosodic features for each word and then calculates the mean, the maximum, the minimum, and the variance of these features for each speaker, that is, the prosody of the whole speech of a speaker is characterized by a 148-dimensional vector. These features differ in the manner in which the information is combined (Figure 1):

(1) onset;
(2) onset position;
(3) offset;

(4) offset position;
(5) maximum;
(6) position of maximum;
(7) minimum;
(8) position of minimum;
(9) regression line;
(10) mean square error of the regression line.

These features are computed for the fundamental frequency (F_0) and the energy (absolute and normalized). Additional features are obtained from the duration and the length of pauses before and after the respective word. Furthermore jitter, shimmer and the length of voiced (V) and unvoiced (UV) segments are calculated as prosodic features.

3.3. Classification System

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.

—*RR*. The total recognition rate determined as the fraction of correctly identified speakers c divided by the number of speakers n:

$$\mathrm{RR} = \frac{c}{n} \cdot 100\%. \tag{1}$$

The RR reports the overall performance of the classifier. This includes the class distribution of the data, that is, if some classes are more frequent than others, their recognition also has more impact on the RR.

—*ROC* denotes the area under the Receiver-Operating-Characteristic (ROC) curve [Fawcett 2006]. A random classifier yields an area of 0.5 while the perfect classifier would yield an area of 1.0.

As classification system we decided for Ada-Boost [Freund and Schapire 1996] in combination with an LDA-Classifier, as was already successfully applied in Hacker et al. [2007]. Investigation of other classifiers yielded similar results. However, as the scope of this article is the investigation of the automatic evaluation of reading disorders and not the comparison of different classifiers, we chose to report only the results obtained by a single classifier setup that we considered state-of-the-art.

4. RESULTS

In a first experiment we evaluated the classification performance for the individual subtests. As reference the outcome of the perceptual evaluation of the respective subtest was chosen. Hence, the number of children and the number of pathologic cases varies (cf. Table II and III). Prosodic features were only used for the evaluation of the read texts in the SLRT4 and SLRT5 subtests. As the other subtests contain single words only, prosodic features did not contribute to the classification performance.

A summary on the different features is presented in Table IV. It reports the mean and standard deviation of duration, WA, WR, and age for each subtest. For the cases of duration and WA, the normal children and the children with reading disorder are at least one standard deviation apart from each other. For WR and age the distributions overlap mostly.

Figure 2 shows a detailed view of the distribution of duration in the SLRT3 subtest. The figure shows that duration is not normally distributed in the dataset. The overlap

Table IV.
Evaluation results for the different information sources on subtest level. Mean and standard deviation are reported although the data may not be normally distributed. Note that the WA and WR values are computed with respect to the reference texts and not with respect to the manual transcriptions of the spoken utterances. Deviation from the target text yields reduction in these values and occurred in the pathologic data as well as the normal data

		SLRT2	SLRT3	SLRT4+5	SLRT7	SLRT8
normal	duration	21.7 ± 6.2	20.5 ± 7.4	35.8 ± 9.92	47.5 ± 13.9	41.9 ± 12.5
	WA	66.7 ± 16.5	83.2 ± 16.6	59.6 ± 15.2	49.1 ± 29.3	50.0 ± 23.2
	WR	70.3 ± 15.5	93.6 ± 8.9	68.9 ± 12.0	76.4 ± 16.0	72.4 ± 13.9
	age	9.6 ± 0.9	9.7 ± 0.6	9.6 ± 1.3	9.5 ± 1.0	9.6 ± 0.9
pathologic	duration	61.4 ± 32.9	61.7 ± 33.4	80.1 ± 38.4	94.9 ± 57.8	84.2 ± 51.1
	WA	25.5 ± 53.1	-7.3 ± 78.6	23.8 ± 45.1	-28.9 ± 79.2	-18.8 ± 76.8
	WR	70.1 ± 16.9	89.6 ± 10.7	71.7 ± 14.7	64.5 ± 17.4	62.4 ± 17.7
	age	9.3 ± 1.1	9.7 ± 0.6	9.6 ± 1.0	9.5 ± 0.9	9.5 ± 1.0

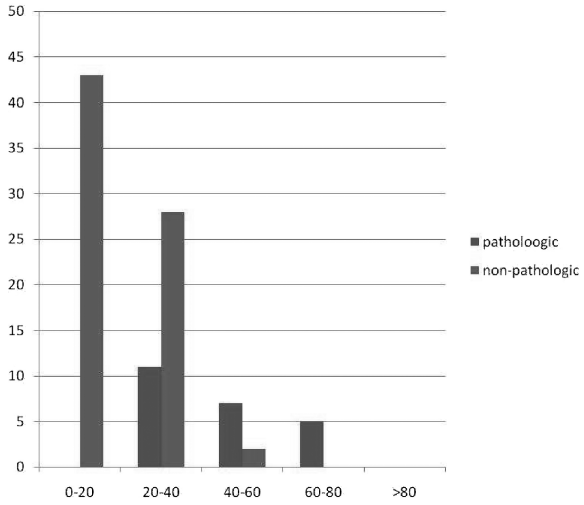

Fig. 2. Distribution of the total reading durations for the normal and pathologic children in the SLRT3 subtest. The overlap of the values is far greater than the mean and standard deviations let expect (cf. Table IV). Both variables are clearly not normally distributed.

between the distributions is higher than the mean and standard deviation in Table IV indicate.

Table V shows the results for the different subtests. In the SLRT2 and SLRT3 subtests duration is the dominant factor. Yet there are cases where additional consideration of the word accuracy helps to increase the classification performance. In all sub-tests the composite classifier shows a significant improvement to the best individual feature ($p < 0.01$). The worst recognition rates appear in the pseudoword subtests. Still the composite classification improves the classification. The SLRT8 subtest has a composite classification rate of 94.8%.

In order to evaluate the overall recognition rate that can be achieved, we use the posterior probabilities, that is, the output probability of the classifiers, for each subtest as input features for a second LDA-Classifier. Again, we use a LOO-setup for the classification experiment. This time the 38 children with reading pathology are to be

Table V.
Overview on the classification results for the different sub-tests. RR is the absolute recognition rate and ROC the area under the ROC curve

	feature	RR [%]	ROC
SLRT2	duration	95.0	0.982
	WA	86.7	0.891
	age	70.6	0.516
	composite	99.2	0.998
SLRT3	duration	89.3	0.955
	WA	86.0	0.900
	age	72.0	0.431
	composite	98.0	0.984
SLRT4+5	duration	82.8	0.917
	WA	82.2	0.836
	age	61.7	0.398
	prosody	76.7	0.849
	composite	95.0	0.980
SLRT7	duration	81.1	0.773
	WA	78.4	0.807
	age	64.4	0.409
	composite	96.7	0.995
SLRT8	duration	80.7	0.775
	WA	80.1	0.851
	age	63.7	0.472
	composite	94.8	0.959

identified. In total 36 of the 38 children could be identified by the automatic system. None of the 82 control children were classified as pathologic. This results in a RR of 98.3% and an area under the ROC curve of 0.958. Figure 3 displays the ROC curve.

5. DISCUSSION

The scope of this article was the automatic detection and classification of reading disorders in children. Therefore, we chose a clinical standard test and recorded the speech of 120 children.

In order to diagnose a reading disorder, the duration of the test has to be investigated and the number of reading errors has to be determined because both variables are related. This was performed according to the manual of the SLRT test. Thirty-eight of the children were diagnosed a reading disorder by our clinical partner. Another 82 did not have a reading disorder.

Table IV summarized the mean values and standard deviations of figures such as the test duration for the pathologic children and the children with normal reading ability. On average, the children with reading disorder take a much more time than the normal children. Hence, time is an important factor to distinguish both groups. However, as shown in Figure 2, the overlap in terms of reading durations is higher than implied by the mean and standard deviations shown in Table IV. The duration is clearly not normally distributed and there is significant overlap, which has to be resolved by another information source.

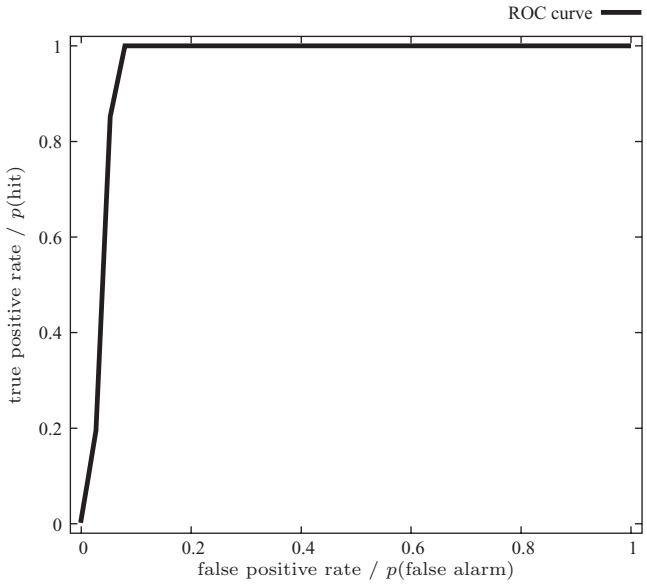

Fig. 3. ROC evaluation of the complete SLRT test: The area under the curve is 0.958.

One feature that is suited for this purpose is the WA. As seen in Table IV, the normal children and the pathologic group show considerable differences. Note that the numbers presented in Table IV were calculated using the target text as reference. Hence, any deviation from the target text is considered as error in the computation of the WA even if it is caused by the reader and not the system. This results in far lower values for the recognition system than the computation of the WA based on the manual transcription of the spoken utterances would produce. However, the transcription would mean the need of a human being who writes down every word that was said. As we aim to develop a fully automatic system, the human transcription was not acceptable for us. The WA even reaches negative mean values for some tests in the pathologic children as it is defined on a scale from minus infinity to 100%. This is related to a frequent number of self repetitions interpreted as wrongly inserted words, and a long duration of the test. If the number of words that were uttered is twice as high as the number of words in the reference, the WA cannot exceed 0% just by its definition. If the number of uttered words is higher, it drops further. Yet, the absolute value of the WA is not important concerning its information content as can be clearly observed in Table V. Word accuracy is an important feature to determine a possible reading disorder of a child.

WR as reported in Table IV is a much better feature to determine the recognition accuracy of the system. In the WR additional words are not regarded as errors. It only comprises substitutions and missing words that would also be denoted as errors by a human observer. In the control group, these values should reach values close to 100% if the evaluation would be performed by a human observer. Yet, the values reach only 90% to 70% as commonly observed in unigram-based speech recognizers for similar tasks. Another interesting property of Table IV is that there is almost no difference in terms of mean value of WR between control group and pathologic children. We relate this to the high number of repetitions and self-corrections in the pathologic children. If each word is recognized at about 70% and repeated twice, the important differences between both groups are almost averaged out in terms of WR.

The age of the children in the normal and the pathologic group is also reported in Table IV. There is no significant difference between both groups in terms of age. Hence, age alone yields very low classification results. In most experiments, the area under the ROC curve is close to 0.5, that is, group assignment is almost random. In some cases, age performs worse than random assignment. This also shows that the sets were well balanced with respect to age.

For the single word subtests (SLRT2, SLRT3, SLRT7, and SLRT8), using the test duration, the WA, and the age-dependent limits of the test, the automatic system could already determine whether the child has a reading disorder. The composite recognition rates lie between 94.8 and 99.2%. The best recognition rates are obtained in the single word reading tests.

Both the SLRT2 and the SLRT3 subtests contain rather simple words which are very common in the German language. The complexity of both subtests is rather low. Only duration already yields high classification rates of up to 95.0% for the SLRT2 sub-test. In the SLRT3 sub-test the performance of duration already decreases to 89.3%. While the classification rate of WA remains in the ball park of 80%, joint classification of all features yields very high recognition rates.

The SLRT4 and SLRT5 subtests present connected text as test material, we regard this test as more complex than the SLRT2 and SLRT3 sub-tests. Time alone yields only 82.8% recognition rate. In these subtests, word recognition becomes slightly less reliable (cf. Table IV), as connected speech is more difficult to recognize. The classification performance using WA only drops to 82.2%. As it is a sentence-based speech test, it is the only subtest in which it makes sense to use prosodic features. The characteristic pauses that occur between the words in children with reading disorders are modeled well by the prosody module. Using only prosody 76.7% of the children are classified correctly.

The SLRT7 and SLRT8 subtests present pseudowords as test material. We regard these two tests as the most complex and difficult tests for the children. In both tests, duration shows about 80% recognition rate. The recognition is worst for these two subtests. As it is a single word test, prosody is of little use. Another problem that occurs here is caused by the structure of the pseudowords: As these words are madeup, they were never seen in the training data of the ASR system. Hence, our polyphone models degenerate. In the SLRT7 data this is not as severe as in the SLRT8 data. As the SLRT7 target words consist of consonant-vowel clusters the polyphones degenerate to only bi- or triphone models in most cases. In the SLRT8 data the pseudowords were built to resemble real words. In most cases, words of the SLRT2 test (mono- and bisyllabic words) were taken and one to two letters were replaced. This results in a very irregular phonemic structure with respect to the German language. In many cases, our polyphone models degrade to monophone models. As a result the recognition performance of the ASR system drops. The degradation of the performance of the classification system, however, is only limited. In the composite classification, both tests achieve 96.7% and 94.8% which is still high, but the lowest classification rate for the subtest level.

In general, the importance of duration as a feature decreases with the complexity of the subtest. The highest rates were observed in the most simple tests and its performance dropped in the more complex tests. It seems that children shift their compensation strategies with the complexity of the test. If the test material is rather simple, the children try to repeat the word until they pronounce it correctly. If the testing material is more complex they do not even realize their own mistake. This is extreme in the pseudowords. Here the words are nonsense words that are unfamiliar to the child. Hence, the test breaks down to a character identification and grapheme-to-phoneme mapping task that is ill posed, as there is often no unique solution. Prior knowledge of the child on existing words does not help. Wrong pronunciations remain

uncorrected. This effect can also be observed in Table IV. While the mean and standard deviation of WR are almost the same for the sub-tests 2 to 5, SLRT7 and SLRT8 show differences in the mean of the WR of about 10%.

Yet, one has to keep in mind that duration is one of the two inherent features to determine the reading pathology. The other inherent feature of the test is reading accuracy. As the results show, only the combination of both yield a consistent high automatic evaluation. As expected, age alone is not a predictor of reading disorders. However, in combination with the other features and the age-dependent limits it helps to improve the classification. Prosody is only helpful in connected read speech.

Furthermore, we investigated whether these classification rates were already enough to determine a reading pathology automatically. We used the posterior probabilities of the first experiment to train another classification system. A classification rate of 98.3% was achieved. Only two children were misclassified. Hence, we could successfully combine the sublevel classifiers to a test-level classifier. In overall recognition rate is higher than in most subtests. Hence, most of the remaining weaknesses of the sub-test classifiers could be corrected by multiple observations of the same child. Note that the final decision on the pathology is based on the result of all subtests. Hence, the training labels of the global task also differ from the labels of all subtests. The evaluation using only SLRT2 would yield 99.2% recognition rate (cf. Table V), but it would only identify 30 of the 38 children with reading disorder (cf. Table III).

In the future this procedure will help in the diagnosis of reading disorders in children as the system can also be used by lay persons with only little understanding of reading disorders. This will simplify the clinical routine. Furthermore, screening of reading disorders is also within the reach of the proposed system. In this manner, many school children could be tested regularly for early intervention.

6. SUMMARY

In this article we present an automatic approach for the classification of reading disorders based on automatic speech recognition and a prosody module. The evaluation is performed on a standardized German reading capability test. To our knowledge such a system has not been published before. The system is web-based and can be accessed from any PC that is connected to the Internet.

Using a database with 120 children, a classification rate of 98.3% could be achieved. The system is suitable for the automatic classification of reading disorders in clinical practice.

REFERENCES

AIST, G. 2000. Helping children learn vocabulary during computer-assisted oral reading. Ph.D. thesis, Carnegie Mellon University.

AIST, G. AND MOSTOW, J. 2009. Designing spoken tutorial dialogue with children to elicit predictable but educationally valuable responses. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*. 588–591.

BATLINER, A., BUCKOW, A., NIEMANN, H., NÖTH, E., AND WARNKE, V. 2000. The Prosody Module. In *Verbal Foundation of Speech-to-Speech Translation*, 106–121.

BEITCHMAN, J. H. AND YOUNG, A. R. 1997. Learning disorders with a special emphasis on reading disorders: A review of the past 10 years. *J. Am. Acad. Child Adolesc. Psych. 36,* 8, 1020–1032.

BLACK, M., TEPPERMAN, J., LEE, S., AND NARAYANAN, S. 2008. Estimation of children's reading ability by fusion of automatic pronunciation verification and fluency detection. In *Proceedings of the 11th International Conference on Spoken Language Processing*. 2779–2782.

DENNIS, I. AND EVANS, J. S. B. T. 1996. The speed-error trade-off problem in psychometric testing. *Brit. J. Psych. 87*, 105–129.

DUCHATEAU, J., CLEUREN, L., HAMME, H. V., AND GHESQUIERE, P. 2007. Automatic assessment of children's reading level. In *Proceedings of the 10th European Conference on Spoken Language Processing*. 1210–1213.

FAWCETT, A. 2006. An introduction to ROC analysis. *Patt. Recog. Lett. 27*, 861–874.

FREUND, Y. AND SCHAPIRE, R. E. 1996. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*. Morgan Kaufmann, 148–156.

GALES, M., PYE, D., AND WOODLAND, P. 1996. Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation. In *Proceedings of the International Conference on Speech Communication and Technology*. ISCA, PA, 1832–1835.

GALLWITZ, F. 2002. *Integrated Stochastic Models for Spontaneous Speech Recognition*. Studien zur Mustererkennung, vol. 6. Logos Verlag, Berlin Germany.

HACKER, C., CINCAREK, T., MAIER, A., HESSLER, A., AND NÖTH, E. 2007. 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)*. Vol. 4., IEEE Computer Society Press, 197–200.

HADERLEIN, T. 2007. *Automatic Evaluation of Tracheoesophageal Substitute Voices*. Studien zur Mustererkennung, vol. 25. Logos Verlag, Berlin, Germany.

HADERLEIN, T., ZORN, D., STEIDL, S., NÖTH, E., SHOZAKAI, M., AND SCHUSTER, M. 2006. Visualization of voice disorders using the sammon transform. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD)*. P. Sojka, I. Kopeček, and K. Pala, Eds., Lecture Notes in Artificial Intelligence, Vol. 4188. Springer, Berlin, 589–596.

KIESSLING, A. 1997. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker, Aachen, Germany.

KITZING, P., MAIER, A., AND AHLANDER, V. L. 2009. Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phoniatrics Vocology 43,* 2, 91–96.

LANDERL, K., WIMMER, H., AND MOSER, E. 1997. *Salzburger Lese- und Rechtschreibtest. Verfahren zur Differentialdiagnose von Störungen des Lesens und des Schreibens für die 1. bis 4. Schulstufe*. Huber, Bern.

MAIER, A., HADERLEIN, T., EYSHOLDT, U., ROSANOWSKI, F., BATLINER, A., SCHUSTER, M., AND NÖTH, E. 2009a. PEAKS—A system for the automatic evaluation of voice and speech disorders. *Speech Comm. 51,* 5, 425–437.

MAIER, A., HADERLEIN, T., AND NÖTH, E. 2006a. Environmental adaptation with a small data set of the target domain. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD)*. P. Sojka, I. Kopeček, and K. Pala, Eds., Lecture Notes in Artificial Intelligence, Vol. 4188. Springer, Berlin, 431–437.

MAIER, A., HADERLEIN, T., STELZLE, F., NKENKE, E. N. E., ROSANOWSKI, F., SCHÜTZENBERGER, A., AND SCHUSTER, M. 2009b. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP J. Audio Speech Music Process*. to appear.

MAIER, A., HÖNIG, F., HACKER, C., SCHUSTER, M., AND NÖTH, E. 2008. Automatic evaluation of characteristic speech disorders in children with cleft lip and palate. In *Proceedings of the 11th International Conference on Spoken Language Processing*. 1757–1760.

MAIER, A., NÖTH, E., BATLINER, A., NKENKE, E., AND SCHUSTER, M. 2006b. Fully automatic assessment of speech of children with cleft lip and palate. *Informatica 30,* 4, 477–482.

MAIER, A., PARCHMANN, C., BOCKLET, T., HÖNIG, F., KRATZ, O., HORNDASCH, S., NÖTH, E., AND MOLL, G. 2009c. On the automatic classification of reading disorders. In *Pattern Recognition in Information Systems*, INSTICC Press, Lisbon, Portugal, 18–28.

MAIER, A., SCHUSTER, M., BATLINER, A., NÖTH, E., AND NKENKE, E. 2007. Automatic scoring of the intelligibility in patients with cancer of the oral cavity. In *Proceedings of the 10th European Conference on Spoken Language Processing*, 1206–1209.

MAIER, A., SCHUSTER, M., EYSHOLDT, U., HADERLEIN, T., CINCAREK, T., STEIDL, S., BATLINER, A., WENHARDT, S., AND NÖTH, E. 2009d. QMOS—A robust visualization method for speaker dependencies with different microphones. *J. Patt. Recog. Resear. 4,* 1, 32–51.

MAIER, A., WENHARDT, S., HADERLEIN, T., SCHUSTER, M., AND NÖTH, E. 2009e. A microphone-independent visualization technique for speech disorders. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*.

MIDDAG, C., MARTENS, J., VAN NUFFELEN, G., AND DE BODT, M. 2009. Automated inteligibility assessment of pathological speech using phonological features. *EURASIP J. Adv. Sign. Process.*

NÖTH, E., BATLINER, A., KIESSLING, A., KOMPE, R., AND NIEMANN, H. 2000. Verbmobil: The use of prosody in the linguistic components of a speech understanding system. *IEEE Trans. Speech Audio Process. 8*, 519–532.

SCHUKAT-TALAMAZZINI, E., NIEMANN, H., ECKERT, W., KUHN, T., AND RIECK, S. 1993. Automatic speech recognition without phonemes. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*. Vol. 1. 129–132.

SCHUSTER, M., HADERLEIN, T., NÖTH, E., LOHSCHELLER, J., EYSHOLDT, U., AND ROSANOWSKI, F. 2006. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *Eur Arch Otorhinolaryngol 263,* 2, 188–193.

SCIPIONI, M., GEROSA, M., GIULIANI, D., NÖTH, E., AND MAIER, A. 2009. Intelligibility assessment in children with cleft lip and palate in italian and german. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*.

STEMMER, G. 2005. *Modeling Variability in Speech Recognition*. Studien zur Mustererkennung, Vol. 19. Logos Verlag, Berlin, Germany.

STEMMER, G., HACKER, C., STEIDL, S., AND NÖTH, E. 2003. Acoustic normalization of children's speech. In *Proceedings of the European Conference on Speech Communication and Technology*. Vol. 2. 1313–1316.

WAHLSTER, W., Ed. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer.

WINDRICH, M., MAIER, A., KOHLER, R., E.NÖTH, NKENKE, E., EYSHOLDT, U., AND SCHUSTER, M. 2008. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. *Folia Phoniatr Logop 60*, 151–156.

Received March 2010; revised October 2010; accepted January 2011