On the Automatic Classification of Reading Disorders

Andreas Maier[1], Caroline Parchmann[2], Tobias Bocklet[1], Florian Hönig[1], Oliver Kratz[2], Stefanie Horndasch[2], Elmar Nöth[1], and Gunther Moll[2]

[1] Lehrstuhl für Mustererkennung (Informatik 5),
Friedrich-Alexander Universität Erlangen Nürnberg,
Martensstr.3, 91058 Erlangen, Germany
[2] Kinder- und Jugendabteilung für Psychische Gesundheit, Universitätsklinikum Erlangen,
Schwabachanlage 6 und 10, 91054 Erlangen, Germany
`Andreas.Maier@cs.fau.de`

Abstract. In this paper, we present an automatic classification approach to identify reading disorders in children. This identification is based on a standardized test. In the original setup the test is performed by a human supervisor who measures the reading duration and notes down all reading errors of the child at the same time. In this manner we recorded tests of 38 children who were suspected to have reading disorders. The data was confronted to an automatic system which employs speech recognition to identify the reading errors. In a subsequent classification experiment — based on the speech recognizer's output and the duration of the test — 94.7 % of the children could be classified correctly.

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 accurate 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 [1]. 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 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-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 [2] 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 [4]) — a client-server-based speech evaluation framework — which was already used to evaluate speech intelligibility in children with cleft lip and palate [5], patients after removal of laryngeal cancer [6], and patients after the removal of oral cancer [7]. 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.

The paper is organized as follows. First the test material, the recorded speech data and its annotation is described and discussed. Next, the automatic evaluation methods, i.e., the speech recognizer and the classifiers, are reported. 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' test material, speech data, and its annotation is described in detail here. Special attention is given to the annotation procedure since the automatic evaluation algorithm aims to be used for clinical diagnosis. Therefore, the annotation should meet clinical standards.

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

2.1 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 sub-tests (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.

We will only report the results obtained for the SLRT7 and SLRT8 sub-tests in the following.

On the one hand, the setup of the perceptual evaluation for all sub-tests is very similar. Therefore, it is not necessary to report the results of all sub-tests. On the other hand, the investigation of pseudo words using automatic systems is described as the most challenging task in the literature [2]. The sub-test SLRT6 also contains pseudo words, but no perceptual evaluations are conducted according to the test manual.

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. 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) 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.

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

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

	SLRT 7		SLRT 8	
grade	# of errors	duration [s]	# of errors	duration [s]
1st	8	144	-	-
2nd	7	113	6	100
3rd	6	78	5	70
4th	5	62	4	55

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 Table 2. All of the children were speculated to have a reading disorder.

2.3 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-tests SLRT7 and SLRT8. In the SLRT7 and the SLRT8 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]. For the case of the SLRT7 data, 12 children exceeded the limit of reading errors while 14 children were above the error limit in the SLRT8 data.

If either of these two boundaries is exceeded by the child, the child is assigned the label "pathologic". In both sub-tests 32 of the 38 children were diagnosed to have pathologic reading.

3 Automatic Evaluation System

The automatic evaluation is based on three information sources:

- The total duration of the test
- The reading error and duration limits (cf. Table 3)
- The word accuracy computed by a speech recognition system

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 — can also easily be obtained (cf. Table 3).

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 Chair of Pattern Recognition (Lehrstuhl für Mustererkennung) of the University of Erlangen-Nuremberg. In this study, the latest version as described in detail in [9] and [10] 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 10 ms. 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 semi-continuous HMMs. The codebook contains 500 full covariance Gaussian densities which are shared by all HMM states. The elementary recognition units are polyphones [11], 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.

We used a unigram language model to weigh the outcome of each word model. It was trained with the reference of the tests. For our purpose it was necessary to emphasize the acoustic features in the decoding process. In [12] a comparison between unigram and zerogram language models was conducted. It was shown that intelligibility can be predicted using word recognition accuracies computed using either zero- or unigram language models. The unigram, however, is computationally more efficient because it can be used to reduce the search space. The use of higher n-gram models was not beneficial.

The result of the recognition is a word lattice. 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 or wrongly inserted words I:

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

Hence, the WA can take values between minus infinity and 100 %.

The speech recognition system had been trained with acoustic information 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 data from 85 male and 47 female adult speakers from all over Germany (2.3 hours of spontaneous speech from the VERBMOBIL project, [13]). The data were recorded with a close-talk microphone with 16 kHz 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 [14]. During training an evaluation set was used that only contained children's speech. MLLR adaptation (cf. [15, 16]) with the patients' test data led to further improvement of the speech recognition system.

3.2 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 three popular measures in order to report the classification accuracy.

- **CL:** The class-wise-averaged recognition rate, or so-called average recall. It is determined as

$$\mathrm{CL} = \frac{1}{K}\left(\sum_{k}^{K} \mathrm{recall}(k)\right) \cdot 100\,\% \tag{1}$$

  where K is the number of classes. The CL is useful if the distribution of the classes is biased.
- **RR:** The total recognition rate determined as the fraction of correctly identified samples c divided by the number of samples n:

$$\mathrm{RR} = \frac{c}{n} \cdot 100\,\% \tag{2}$$

  The RR reports the overall performance of the classifier including the class distribution of the data.
- **ROC** denotes the area under the Receiver-Operating-Characteristic (ROC) curve [17]. 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 [18] in combination with an LDA-Classifier as simple classifier as it was already successfully applied in [19].

4 Results

Table 4 shows the results of the classification experiment "reading error". The task was to determine automatically whether the age-dependent limit of reading errors was exceeded or not. For both sub-tests, the classification performance using duration information and WA only is moderate. If the age-dependent limit which is dependent on the school grade of the child is also provided to the classifier, the performance increases for

Table 4. Overview on the classification results for the task "reading error". CL is the average recall, RR the absolute recognition rate and ROC the area under the ROC curve.

	SLRT 7			SLRT 8		
feature set	CL [%]	RR [%]	ROC	CL [%]	RR [%]	ROC
duration and accuracy	59.5	65.8	0.74	61.6	68.4	**0.74**
+ age-dependent limits	**90.1**	**89.5**	**0.97**	**68.2**	**71.1**	0.62
+ actual age	77.6	81.6	0.81	50.3	57.9	0.56

Table 5. Overview on the classification results for the task "pathologic". CL is the average recall, RR the absolute recognition rate and ROC the area under the ROC curve.

	SLRT 7			SLRT 8		
feature set	CL [%]	RR [%]	ROC	CL [%]	RR [%]	ROC
duration and accuracy	**90.1**	**94.7**	**0.99**	68.7	81.6	0.66
+ age-dependent limits	83.3	84.7	**0.99**	70.3	84.2	0.76
+ actual age	81.8	91.1	**0.99**	**88.5**	**92.1**	**0.83**

both sub-tests (90.1 % CL for SLRT7 and 68.2 % CL for SLRT8). The actual age — defined by the date of birth of the child and the date and time of recording — did not yield any improvement for this classification task.

As a second experiment, the use of the classification system to automatically determine reading disorders was investigated. Now, the task was to classify whether the child has a reading disorder or not. Table 5 reports the results. Using only duration and WA, high recognition rates can already be obtained for the SLRT7 sub-test. For the case of the SLRT8 sub-test, more information is required to obtain such high classification rates. If the age-dependent limits of the SLRT8 and the actual age of the child are supplied to the classifier, a CL of 88.5 % is achieved.

5 Discussion

The scope of this paper was the automatic detection and classification of reading disorders in children. Therefore, we chose a clinical standard test and recorded 38 children who were speculated to have reading disorders.

In order to diagnose a reading disorder, the time 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.

Next, these data were confronted to an automatic evaluation routine based on an automatic speech recognition system. For the SLRT7 test, using the test duration, the WA, and the age-dependent limits of the test, the automatic system could already determine whether the child exceeded the number of reading errors or not at a CL of 90.1 % (89.5 % RR). For the SLRT8, however, only 68.2 % accuracy were achieved. Figure 1 shows the relation between WA and the duration for the SLRT8 sub-test. Although both are correlated, duration and WA can be used to distinguish children with many reading errors from children with only few reading errors. However, both clusters are scattered

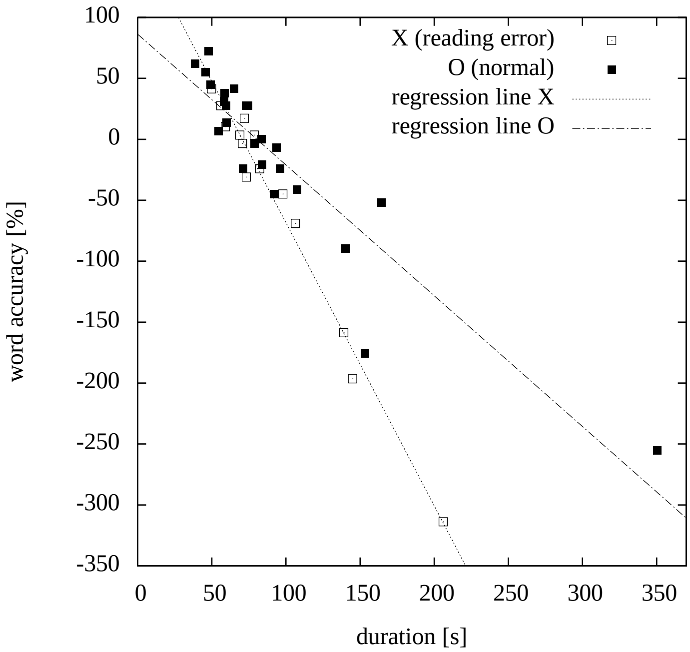

Fig. 1. The plot shows the children of the SLRT8 sub-test: The regression lines show that the additional use of speech recognition helps to differentiate between the children with many reading errors and the ones with few reading errors.

into each other. If the reading error limit is also supplied the performance of the classification system increases. The actual age of the children did not contribute. This may be related to the boosting algorithm. It emphazises the wrong features and classifies according to the age instead of the grade.

In a second experiment we investigated whether these classification rates were already enough to determine a reading pathology automatically. In the SLRT7 sub-test, 90.1 % CL (94.7 % RR) were achieved. Investigation of the SLRT8 sub-test showed that a high classification rate of 88.5 % CL (92.1 % RR) could also be achieved. Hence, the proposed system is suitable for the automatic classification of reading disorders, even though the classification of the reading errors is not perfect. Figure 2 shows this process: If duration and reading errors are taken into account the group of non-pathologic children can be found on the top left of the plot. However, both groups still overlap. Further prior information helps to distinguish them from each other. In this experiment the limits and the true age of the child improve the classification further. In further experiments using the other sub-tests and a control group of non-pathologic school children, we will investigate this effect in more detail.

In the future this procedure will help in the diagnosis of reading disorders in children. The system can also be used by lay persons with only little understanding of reading disorders. Screening of reading disorders is also within the reach of the proposed system.

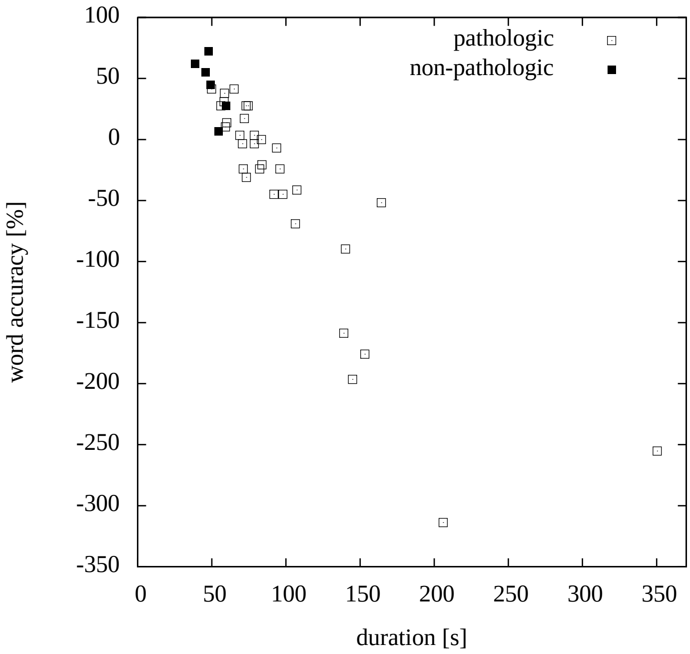

Fig. 2. The plot shows the children of the SLRT8 sub-test: The non-pathologic children can be found in the upper left corner of the diagram. The data points of the pathologic children are scattered to the bottom right. Duration and word accuracy (WA) alone are not sufficient to separate both clusters.

6 Summary

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

Using a database with 38 children classification rates of up to 94.7 % (RR) could be achieved. The system is suitable for the automatic classification of reading disorders.

References

1. I. Dennis and J. St. B. T. Evans, "The speed-error trade-off problem in psychometric testing," *British Journal of Psychology*, vol. 87, pp. 105–129, 1996.
2. 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*, 2008, pp. 2779–2782.

3. J. Duchateau, L. Cleuren, H. Van 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*, 2007, pp. 1210–1213.
4. 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*, vol. 51, no. 5, pp. 425–437, 2009.
5. A. Maier, E. Nöth, A. Batliner, E. Nkenke, and M. Schuster, "Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate," *Informatica*, vol. 30, no. 4, pp. 477–482, 2006.
6. M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, and F. Rosanowski, "Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating," *Eur Arch Otorhinolaryngol*, vol. 263, no. 2, pp. 188–193, 2006.
7. 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*, vol. 60, pp. 151–156, 2008.
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. F. Gallwitz, *Integrated Stochastic Models for Spontaneous Speech Recognition*, vol. 6 of *Studien zur Mustererkennung*, Logos Verlag, Berlin (Germany), 2002.
10. G. Stemmer, *Modeling Variability in Speech Recognition*, vol. 19 of *Studien zur Mustererkennung*, Logos Verlag, Berlin (Germany), 2005.
11. E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck, "Automatic Speech Recognition without Phonemes," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Berlin (Germany), 1993, vol. 1, pp. 129–132.
12. K. Riedhammer, G. Stemmer, T. Haderlein, M. Schuster, F. Rosanowski, E. Nöth, and A. Maier, "Towards Robust Automatic Evaluation of Pathologic Telephone Speech," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, Kyoto, Japan, 2007, pp. 717–722, IEEE Computer Society Press.
13. W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Berlin (Germany), 2000.
14. G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic Normalization of Children's Speech," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, 2003, vol. 2, pp. 1313–1316.
15. M. Gales, D. Pye, and P. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," in *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, Philadelphia, USA, 1996, vol. 3, pp. 1832–1835, ISCA.
16. A. Maier, T. Haderlein, and E. Nöth, "Environmental Adaptation with a Small Data Set of the Target Domain," in *9th International Conf. on Text, Speech and Dialogue (TSD)*, P. Sojka, I. Kopeček, and K. Pala, Eds., Berlin, Heidelberg, New York, 2006, vol. 4188 of *Lecture Notes in Artificial Intelligence*, pp. 431–437, Springer.
17. A. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.
18. Yoav Freund and Robert E. Schapire, "Experiments with a new boosting algorithm," in *Thirteenth International Conference on Machine Learning*, San Francisco, 1996, pp. 148–156, Morgan Kaufmann.
19. 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)*, Hawaii, USA, 2007, vol. 4, pp. 197–200, IEEE Computer Society Press.