From G. Kokkinakis et al., *10th International Conference on SPEECH and COMPUTER (SPECOM 2005)*, pp. 123–126,
© 2005 University of Patras, Moskow State Linguistics University, ISBN 5-7452-0110

Assessment of Non-Native Children's Pronunciation: Human Marking and Automatic Scoring

*Christian Hacker, Anton Batliner, Stefan Steidl, Elmar Nöth, Heinrich Niemann, Tobias Cincarek**

Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany
hacker@informatik.uni-erlangen.de

Abstract

The paper investigates automatic rating of non-native children's pronunciation. We have designed a set of 28 pronunciation-features; when classification is performed in high-dimensional feature space best recognition-results can be achieved. Different measures to evaluate inter-rater agreement and the machine score are proposed. In the European project Pf-Star data of native and non-native children has been recorded; the German children reading English texts have been graded by 13–14 raters. When classifying 5 sentence-level marks the result can be interpreted as 73 % correct. Looking at a longer context, recognition becomes more robust. On the speaker level error and correlation is comparable with some of the human raters.

1. Introduction

The development of useful educational software for computer assisted language learning (CALL) requires robust scoring algorithms that rate the student's skills in a similar manner as a human teacher would do. In this paper we focus on an application that supports second-language learning for children. The first challenge is to automatically recognize children's speech; this has been investigated in the European project Pf-Star (http://pfstar.itc.it/). Caused by higher spectral variability or higher variability in speaking rate, vocal effort, and degree of spontaneity, it is more difficult to recognize young speakers than adults [1, 2]. Another challenge is the recognition and finally the rating of non-native speech which we will focus on in the following. All Pf-Star partners recorded non-native children; for comparison, our partners in Birmingham collected similar native data. The pronunciation of the German children reading English sentences has been marked by several teachers of English. In this paper we will compare the teachers' ratings and measure the reliability. For other databases human ratings have been compared e.g. in [3] or [4].

Research on automatic scoring of non-natives' pronunciation has been carried out on the phone-level e.g. in [5]. The GOP-measurement (*Goodness of Pronunciation*) calculates the posterior probabilities of the desired phone; for this purpose forced alignment scores and the output of a phone-recognizer are compared. Phone-level scoring is necessary to localize the pronunciation error. However, the system will not be accepted by the user if too many false alarms occur and the user gets frustrated. Since particularly the pronunciation of beginners is rather poor, an additional measurement is required to reliably judge the candidate's English on a higher level (sentence- or text-level) based on a longer observation. Neumeyer et al. [3] automatically score non-natives on sentence and speaker-level. The inter-rater open-correlation is 0.78 (sentence) and 0.87 (speaker). Correlations are calculated with machine scores obtained from different features: HMM log-likelihood scores, posterior probabilities of the desired phone for each phone-segment, word or phone recognition rate, duration, and syllabic timing. Different combination techniques of sentence based scores are investigated in [6]: with neural networks a correlation of 0.64 is achieved. Different aspects of human ratings are compared with several machine scores for sentences in [4].

In the following different pronunciation-features are combined and extended to an 28-dimensional feature vector. Machine scores are computed with the LDA-classifier. After a description of the data we will propose measures for the agreement among teachers and between the human and automatic ratings. Then the teacher's agreement will be analyzed. In the last part the automatic scoring of utterances, texts and speakers is described and discussed.

2. Corpora

The PF-STAR NON-NATIVE-database contains recordings of German children reading English texts: 57 children (26m, 31f), age 10 – 15, from two different schools (OHM and MONT[1]). Altogether the database comprises 3.4 hours of speech (4627 utterances). Most children had been learning English for half a year only. They were reading known texts from their text book and some phrases and single words, which have been recorded by all partners in the Pf-Star project. The recordings include

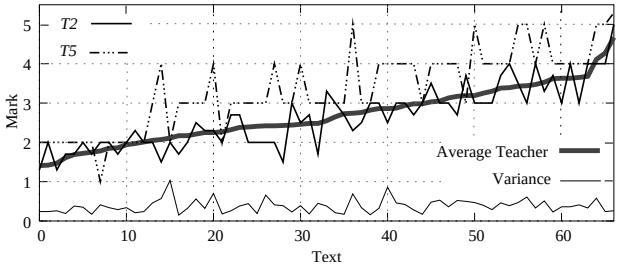

Figure 1: *The mean marks over all teachers and marks of teachers T2 and T5. How can we measure the agreement?*

reading errors, repetitions of words, word fragments, and non-verbals. The total size of the vocabulary is 940 words. The PF-STAR NATIVE-corpus (14.2 hours, 1740 words) contains British children recorded by the University of Birmingham [7]: 159 children, age 4 – 14. In the following the NATIVE data is

* Now with Graduate School of Information Science, Nara Institute of Science and Technology, Japan

[1] Ohm-Gymnasium and Montessori Schule, Erlangen

	text-level			speaker-level		
	min	mean	max	min	mean	max
C	0.62	0.78	0.89	0.63	0.80	0.94
E	13.8	20.1	31.5	9.4	17.4	32.3
E_{norm}	13.5	17.1	21.9	9.3	13.8	21.4
CL3	39.9	72.4	89.7	52.8	77.8	93.8
CL5	31.5	51.4	67.9	27.8	45.4	70.4

Table 1: *Agreement of human ratings: Open-correlation C; open-error E and E_{norm} (same mean for all raters) in %. In the discrete case for 3 and 5 classes:* CL3 *and* CL5 *in %*

used for the training of the HMMs of the recognizer and for some comparison in feature space.

The NON-NATIVE data has been graded with 1 (best) to 5 (worst pronunciation). A German student of English (graduate level) has rated all the data on the utterance-level (rater *S*). Most of the sentences are rated with 2 (50 %, 1: 16 % , 3: 20 % , 4: 8 % , 5: 5 %). Recordings from the OHM school (32 pupils) have been annotated additionally by 12 teachers (raters *T1 – T12*) on the text-level and by a native teacher (rater *N*). *T1 – T7* have many years of teaching experience, the other raters less than two years. 4 of the experienced teachers were asked to rate the data half a year later again. For the speaker-level, ratings are obtained by averaging the text-ratings. An additional speaker rating is performed by rater *S*. Each speaker read 3.8 texts in the average, each text contains around 11 utterances.

3. Comparison of ratings

The *correlation* $\mathrm{Cor}(\boldsymbol{x}^i, \boldsymbol{x}^j)$ is the measurement most frequently used when ratings are compared in a metric space $\mathcal{M}$. $\boldsymbol{x}^k \in \mathcal{M}^N$ are vectors, where the N components are the labels of rater k. In our case the domain is $\mathcal{M} = [1, 5]$ which includes the discrete marks $1 - 5$. Also intermediate marks (e.g. 1.3, 1.5, 1.7 and so on) were allowed and used by some teachers very frequently but by others very rarely. For K raters the *open-corelation* of rater k is defined as

$$C(\boldsymbol{x}^k) = \mathrm{Cor}(\boldsymbol{x}^k, \frac{1}{K-1}\sum_{i \neq k} \boldsymbol{x}^i) \tag{1}$$

In this paper we compare the open-correlation of human raters with the correlation between the machine score and the average rater. Fig. 1 shows the ratings of *T2* ($C(\boldsymbol{x}^{\mathrm{T2}}) = 0.86$) and *T5* ($C(\boldsymbol{x}^{\mathrm{T5}}) = 0.79$). Both ratings are highly correlated, however, *T5* gives systematically higher marks, which cannot be measured with the correlation. Thus we introduce the *error* of rater i (vector components $x_n^k, n = 1, \ldots, N$)

$$\mathrm{Err}(\boldsymbol{x}^i, \boldsymbol{x}^j) = \frac{1}{N}\sum_n^N \frac{|x_n^i - x_n^j|}{\max(x_n^j - m_{\min}, m_{\max} - x_n^j)} \tag{2}$$

where the denominator calculates the maximal possible error by comparing the reference rater j with the minimal ($m_{\min} = 1$) and maximal ($m_{\max} = 5$) mark $m \in \mathcal{M}$. The error-measure is introduced in [8] but normalized with $m_{\max} - m_{\min}$. The *open-error* $E(\boldsymbol{x}^k)$ is calculated analogous to Eq.1. The ratings in Fig. 1 are assessed with $E(\boldsymbol{x}^{\mathrm{T2}}) = 12\,\%$ and $E(\boldsymbol{x}^{\mathrm{T5}}) = 24\,\%$.

A measurement for discrete automatic scoring is the *recognition rate* RR. In the following an items (sentence, text) is considered as correctly *recognized* or *classified*, if the decoder (machine or human) agrees with the reference (human). To guarantee robust recognition for all classes on unbalanced data, the *class-wise averaged* recognition rate CL is used, which is the unweighted average recall. RR5 and CL5 give classification rates for the 5-class problem. The coarser 3-class problem (RR3, CL3) only discriminates the classes *good pronunciation* (marks 1, 2), *bad pronunciation* (marks 3, 4), and *mispronounced* (mark 5). For the human ratings we calculate analogous to Eq.1 the open-CL: one rater is the decoder whereas the mean of all other raters is the reference.

Other measures for the agreement of ratings are e.g. the weighted κ [9]. Here, only $\kappa = 0.38$ is achieved since many raters do not use the whole range of $\mathcal{M}$. $0.1 < \kappa \leq 0.4$ indicates weak agreement. For non-metric ratings we proposed a measure in [10].

4. Human ratings

For the 12 teachers and the native rater we calculated C, E , E_{norm}, CL3, and CL5 ($K = 13$ in Eq.1). Values for these measures (maximum, minimum and mean over all raters) are given in Tab. 1; the mean open-correlation of all text-level ratings is 0.78. The mean open-error is 20 %, after normalization in a way, that the mean of each teacher's ratings equals 3 (E_{norm}), the error decreases to 17 %. For the discrete classification with 3 classes in the average 72 % CL3 is achieved and 51 % CL5. As 4 teachers judged the data 2 times, we expect the maximal possible agreement if we calculate correlation and error for these pairs. The intra-speaker correlation is between 0.75 and 0.83 (mean: 0.81), the error E is between 15 % and 17 % (mean: 16 %). In the average, better values are obtained for the intra-speaker correlation and error than for the mean open-correlation and -error (Tab. 1).

On the speaker-level we have $K = 14$ raters (additionally rater *S*). As can be seen in Tab. 1, the mean open-correlation is 0.80, the lowest correlation is observed for teacher *T9* ($C = 0.63$) and rater *S* ($C = 0.64$), the highest error for *S* ($E_{\mathrm{norm}} = 21\,\%$) and for the native teacher ($E_{\mathrm{norm}} = 17\,\%$, $C = 0.75$). To be native does not mean better agreement with the average teachers: in our case the native teacher has less teaching experience and less practice in grading German children.

5. Automatic scoring

For the extraction of pronunciation features word- and phone-recognizers are required. We used the HTK toolkit to estimate monophone HMMs and language models (LMs) with the PF-STAR NATIVE data. In order to test with non-natives, LMs are estimated on the NON-NATIVE corpus. LMs are build from the original texts, since they can be assumed to be known in our scenario; the perplexity of both unigram LMs is around 150. The recognizers are based on 12 Mel-cepstrum features, the energy, and 13 first and 13 second order derivatives. With a unigram LM 19.6 % word accuracy (WA) are achieved for the NATIVE test data and 18.2 % WA for NON-NATIVE (bigram: 42.8 vs. 38.1 % WA, perplexity 34 vs. 17). The low recognition rates show the difficulties caused by the wide age-range of the native children and the dissimilar pronunciation by German children with only little practice in English.

The recognized phone sequences together with the likelihood scores as well as the forced alignment of the data together with scores are the basis for the pronunciation feature extraction. In addition, a phone bigram model and a duration look-up-table (D-LUT) with native phone-duration statistics are es-

	C	E	CL3	CL5
Rate-Of-Speech	**0.35**	41.2	41.2	28.6
Pauses	0.31	41.8	41.4	**29.1**
DurationLUT	0.13	42.1	40.0	23.7
DurationScore	**0.35**	**40.6**	**44.9**	28.5
Likelihood	0.30	41.8	41.1	26.4
LikeliRatio	0.21	41.8	**44.0**	27.0
PhoneSeq	**0.34**	41.5	42.4	28.6
Accuracy	0.20	41.7	**47.8**	**29.1**

Table 2: *Pronunciation features (E,* CL3 *and* CL5 *in %).*

timated on the TIMIT[2] database. A set of 28 features has been designed for utterance-level scoring. It is an extension of the features proposed in [3]. For details, please refer to [11]. Features within one of the following eight groups (number given in brackets) differ basically in the way of normalization:

Rate-Of-Speech (ROS) (5): # of phones or words per sentence, both reciprocals and the proportion of phonation time.

Pauses (2): Duration of between-word pauses, number of pauses longer than 0.2 sec.

DurationLUT (2): The actual duration of phones is compared with the expected duration from the D-LUT. As features we take the mean duration deviation and the scatter.

DurationScore (2): $P(t|p,\boldsymbol{o})$ is the probability of the observed duration t (normalized with the ROS) given the desired phone p and the acoustic observation $\boldsymbol{o}$. The duration distribution is estimated on native data (TIMIT). We sum up $P(t|p,\boldsymbol{o})$ over all phones of an utterance and normalize e.g with the # of phones.

Likelihood (9): This features are based on the log-likelihood $\log P(\boldsymbol{o}|\lambda_q)$ of the acoustic observation $\boldsymbol{o}$ given the HMM λ of the recognized phone q. For some components the likelihoods are normalized by the phone-duration or the expected duration from the D-LUT. After that, the log-likelihoods are summed up over all frames and normalized e.g. with the ROS

LikeliRatio (3): $\log P(\boldsymbol{o}|\lambda_p) - \log P(\boldsymbol{o}|\lambda_q)$ is the ratio between the likelihood obtained by the forced alignment and the one obtained by phone recognition. We normalize with the number of phones, the ROS or the expected duration.

PhoneSeq (3): The probability of the recognized phone sequence given a phone bigram LM. Again we normalize with the # of phones or the ROS or take the not normalized values.

Additionally word- and phone-**Accuracy** are used as features. Tab. 2 summarizes the feature groups and shows the optimal values for C, E, CL3, and CL5 per feature-group . Details of the experimental setup are given in the next section.

6. Experimental results

The experimental setup is the following: first pronunciation features are calculated for the NON-NATIVE database. We utilize forced alignment and recognition results based on HMMs trained on native data. Decorrelation and feature reduction of the 28-dimensional feature vectors is performed with principal component analysis (PCA). With the resulting features LDA classifiers are trained using the leave-one-speaker-out (*loo*) technique. The output of all test iterations is accumulated and afterwards evaluated. Dependent on the training (whether we train 5 classes for marks 1–5 or 3 classes for marks 1, 3, 5; mapping as described in Sec.3) we get for each test-utterance 5 or 3 posterior probabilities: $P(i|\boldsymbol{u})$ is the probability of mark i given the utterance-level features $\boldsymbol{u}$.

[2]http://www.ldc.upenn.edu/ catalog number LDC93S1

features	PCA	C	E	CL3	CL5
all	$28 \rightarrow 28$	**0.33**	**29.2**	49.0	33.3
all	$28 \rightarrow 14$	0.32	29.8	**50.2**	**33.4**
3 (best C)	$3 \rightarrow 3$	0.24	32.3	41.3	27.8
3 (best CL3)	$3 \rightarrow 3$	0.21	30.6	48.6	30.3

Table 3: *Automatic classification after feature reduction/ decorrelation with PCA (E,* CL3, CL5 *in %)*

Discrete classification results are obtained if we decide for the mark with maximum $P(i|\boldsymbol{u})$. These results are evaluated with CL3 and CL5. In order to measure C and E we use a continuous classification score: the expectation $\sum_i iP(i|\boldsymbol{u})$ is calculated over all marks $i \in \{1,2,3,4,5\}$ of the 5-class problem. On the utterance-level we use the ratings of *S* for training and testing; for final evaluation on the text- and speaker-level the classification result is compared with the average of the teacher's ratings.

Utterance-level scoring. Tab. 3 shows the recognition results for the utterance-level. After reduction of the 28-dimensional feature-vectors to 14 principal-components a more robust training is possible and better recognition rates are achieved. For the discrete 5-class-task up to 33.4 % CL5 (28.7 % RR5[3]) are obtained, for the 3-class-task 50.2 % CL3 (52.6 % RR3[4]). Rates for the 5-class-task are lower, which is caused by a clear overlap of neighboring classes. If a confusion of neighboring marks is tolerated, the result can be interpreted as 72.9 % RR. This phenomenon is also measured by the error E, that is only around 30 % whereas $1 - \text{CL5}$ would be more than 66 %. The correlation is $C = 0.33$ in the best case. In comparison with Tab. 2 correlation decreases but recognition rates are higher. If we combine the best 3 features in terms of correlation from Tab. 2 we get low recognition rates; if we combine the best 3 features in terms of CL3, acceptable recognition rates are achieved. However, with higher-dimensional feature-vectors correlation and recognition rates can be increased.

Next we will look at the distribution in feature space. Fig. 2 (left) shows the distribution of some of the optimal features in Tab. 2 (*LikeliRatio* and *DurationScore*): three classes for the

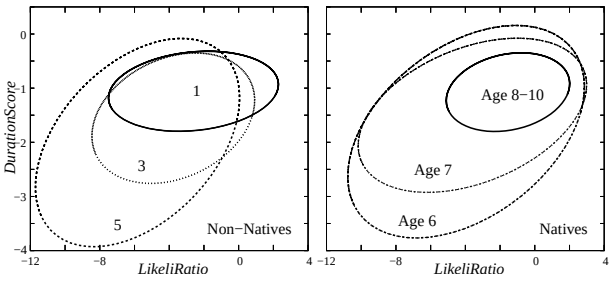

Figure 2: *Pronunciation feature space: clusters for marks 1 – 3 of non-natives (l.) and clusters for age-groups of natives (r.).*

marks 1, 3, and 5. Unfortunately, the classes clearly overlap. This may be caused by both, not enough discriminant features nor precise ratings by only one labeler. However, as can be seen in Fig. 2 (right), the cluster for mark 1 is covered by the class of native children in the age of 8–10, which complies with the assumption that natives would be marked with 1. The different pronunciation of good non-native speakers and natives is com-

[3]Recalls: 1: 55.2 %, 2: 19.5 %, 3: 29.2 %, 4: 24.9 %, 5: 38.4 %

[4]Recalls: 1: 59.2 %, 3: 37.3 %, 5: 54.0 %

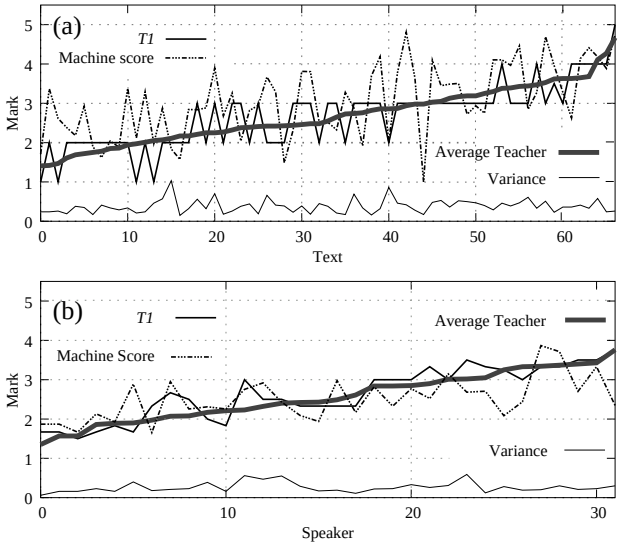

Figure 3: *Text (a) and speaker (b) rating: Mean marks of all teachers, marks of teacher T1, automatic scoring.*

pletely eliminated with our feature set. However, even for natives poor pronunciation can be observed for young children. In the 5-class task natives with age 8–10 are classified with mark 1 (53 %), 2 (17 %) and 3 (18 %) whereas 6–7 year old natives are classified with 1 (49 %), 2 (11 %) and 5 (30 %).

Text-/speaker-level scoring. More precise machine scores are achieved, if we use more information than only one utterance. In the following we calculate a continuous machine score as described above, but from the 3-class task. Then we build the average score over all utterances of a text or a speaker. As for the text based scoring the correlation with the average rater is $C = 0.53$. The confidence interval on a 95 % significance level is $[0.38; 0.65]$. For the error we obtain $\mathrm{E} = 25\,\%$. The text based machine-rating, the average of the teachers' marks, and *T1* ($C = 0.83$, $\mathrm{E} = 15\,\%$) are compared in Fig. 3a. The recognition rate is 59 % CL3 (32 % CL5).

For the speaker based rating the correlation is even 0.61 ($E = 16\,\%$) which is close to rater *T9* ($C = 0.63$, $E = 16\,\%$). In Fig. 3b the automatic score is compared with the average teacher and teacher *T1* ($C = 0.89$, $E = 9\,\%$).

7. Summary

In this paper we presented a high dimensional feature set for the classification of non-natives' pronunciation. The combination of many features based on different normalization techniques increases recognition rates. All experiments are conducted with the PF-STAR NON-NATIVE-corpus with ratings by several teachers. We discussed 4 different measures for the agreement: the correlation, the error that takes into account the distances between marks, and the class-wise averaged recognition rate for the discrete 3 and the 5-class task, that unfortunately punishes confusion of neighboring classes in the same way as confusion of far distant classes. Both kinds of measures should be optimal: the correlation *and* one of the proposed recognition-rates. Sentences have been assessed only by rater *S*, thus no robust reference was available. 3 classes are recognized with 50 % CL3, for 5 classes, only every third utterance is correctly classified. However, since especially neighboring classes overlap, less than one third of the maximal possible deviation from the reference mark occurs (error 30 %); the correlation is 0.33. Between teachers the correlation is 0.8 on text and speaker-level. With machine scores we achieve 0.5 on the text-level and 0.6 on the speaker-level. This automatic result is in the same rage as rater *S* or *T9*. The error drops on the speaker-level to 16 % while human raters make 14 % errors in comparison with the average teacher. When testing natives with our classifier 70 % could be classified as good speakers (mark 1 and 2), in particular the 8–10 year old children.

8. Acknowledgments

A part of this work was funded by the European Commission (IST programme) in the framework of PF-STAR (Grant IST-2001-37599) and by the German Federal Ministry of Education and Research (BMBF) in the frame of SmartWeb (Grant 01 IMD 01 F). The responsibility for the content lies with the authors.

9. References

[1] A. Potamianos and S. Narayanan, "Robust Recognition of Children's Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603– 616, 2003.

[2] G. Stemmer, "Modeling Variability in Speech Recognition," Ph.D. dissertation, Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Germany, 2005.

[3] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality," *Speech Communication*, vol. 30, pp. 83–93, 2000.

[4] C. Cucchiarini, H. Strik, and L. Boves, "Different Aspects of Expert Pronunciation Quality Ratings and Their Relation to Scores Produced by Speech Recognition Algorithms," *Speech Comm.*, vol. 30, pp. 109–119, 2000.

[5] S. M. Witt and S. J. Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning," *Speech Communication*, vol. 30, pp. 95–108, 2000.

[6] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of Machine Scores for Automatic Grading of Pronunciation Quality," *Speech Communication*, vol. 30, pp. 121–130, 2000.

[7] S. D'Arcy, L. Wong, and M. Russell, "Recognition of Read and Spontaneous Children's Speech Using two new Corpora," in *Proc. ICSLP*, Korea, 2004.

[8] C. Teixeira, F. Horacio, E. Shriberg, K. Precoda, and K. Snmez, "Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language learners," in *Proc. ICSLP*, 2000, pp. 187–190.

[9] F. Krummenauer, "Erweiterungen von Cohen's kappa-Maß für Multi-Rater-Studien: Eine Übersicht," *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, vol. 30, pp. 3–20, 1999.

[10] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "'Of All Things the Measure is Man' – Automatic Classification of Emotions and Inter-Labeler Consisency," in *Proc. ICASSP*, vol. I, 2005, pp. 317–320.

[11] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann, "Pronunciation Feature Extraction," in *Pattern Recognition, 27th DAGM Symposium,Vienna, Austria*. Berlin, Heidelberg: Springer, 2005, to appear.