PROC. ICASSP 2007, VOLUME 4, PP. 197-200, HONOLULU, HAWAII, APRIL 15-20, 2007.
©2007 IEEE, ISBN: 1-4244-0728-1 (DVD)

BOOSTING OF PROSODIC AND PRONUNCIATION FEATURES TO DETECT MISPRONUNCIATIONS OF NON-NATIVE CHILDREN

Christian Hacker[1], *Tobias Cincarek*[2], *Andreas Maier*[1], *André Heßler*[1], *Elmar Nöth*[1]

[1]Chair of Pattern Recognition, University of Erlangen-Nuremberg, Germany
[2]Graduate School of Information Science, NAIST, Nara, Ikoma, Japan
hacker@informatik.uni-erlangen.de

ABSTRACT

Commercial products that support L2-learners with computer assisted pronunciation training usually focus per exercise only on one possible pronunciation mistake that is typical for speakers of the respective L1 group. Acoustic models for words with wrong pronunciation are added to the system. In the present paper a more general approach with features that have proved to be widely independent of the learners' mother tongue is proposed. It is able to take various possible mistakes into consideration all at once. High dimensional feature vectors that encode prosodic varieties and differences of reference and recognized sentences are analyzed. With the ADABOOST algorithm those features are found, which contain the most important information to assess German children learning English. With 35 features 89 % of the agreement of experts is achieved.

Index Terms— feature extraction, speech intelligibility

1. INTRODUCTION

A lot of research has been focused on computer assisted pronunciation training (CAPT) in the recent years which supports in most cases learners of the second language (L2) English. Primarily, language specific, rule based approaches have been investigated; all rules depend on the learners' mother tongue. For the European languages, an important research project was ISLE (described by Herron et al.[1]), that has focused on adult German and Italian learners. In the European project Pf-Star [2] we focused on speech technologies for children. Based on these technologies, different systems for the assessment of pronunciation have been developed in our institute, e.g to objectively evaluate speech disordered children with a cleft lip and palate. *Caller* is a system for **c**omputer **a**ssisted **l**anguage **l**earning from **Er**langen [3], a client/server system that can be started in a browser; speech is analyzed on a server, e.g. placed in a school's computer room (Fig. 1).

In many common approaches to detect mispronunciations (like in [1]) rules for possible mispronunciations are introduced, e.g. Germans tend to pronounce the "w" in "where" like the "v" in "very". Acoustic models for the wrong pronunciation are added to the speech recognizer. The word with the most likely mispronunciation is found trying forced alignment on different word sequences. To reduce complexity, many systems provide those alternative pronunciations for only one word per sentence. Consequently mispronunciation of all other words is not detected at all. *Caller* integrates a similar approach using speech recognition (Fig. 1, C), that is discussed in [3]. In the "Fluency pronunciation trainer" Eskenazi et al. [4] investigate forced alignment and prosody to pinpoint pronunciation errors. Some research is focused on pronunciation features, e.g. on the phone-level the GOP-measurement (*Goodness of Pronunciation*): Witt et al. [5] calculate the posterior probabilities of all desired phones; for this purpose forced alignment scores and the output of a phone-recognizer are compared. Additional measurement on higher levels (sentence- or text-level) are useful to adapt to the speakers' proficiencies or to calculate a mark or score per exercise. Neumeyer et al. [6] automatically score non-natives on sentence and speaker-level: Correlations are calculated between human experts and different features, e.g. 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 [7]. Different aspects of human ratings are compared with several machine scores for sentences in [8].

In the present paper, pronunciation features and prosodic features are combined to 176 dimensional feature vectors to detect wrongly pronounced words. First, we describe a corpus with German children reading English sentences that has been rated by 10 experts. In the next section the features are described. With the ADABOOST algorithm we find out, which features provide optimal classification results. Finally, experimental results are discussed.

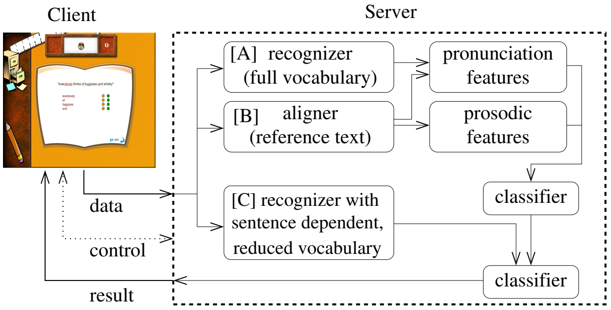
Client
Server
[A] recognizer (full vocabulary)
[B] aligner (reference text)
pronunciation features
prosodic features
[C] recognizer with sentence dependent, reduced vocabulary
classifier
classifier
data
control
result

Fig. 1. The *Caller*-architecture. In the present paper we focus on pronunciation and prosodic features.

2. CORPUS AND EXPERT RATINGS

The PF-STAR NON-NATIVE-database [2] contains recordings of German children reading English texts: 57 children from two different schools OHM and MONT. Altogether the database comprises 3.4 hours of speech (4627 utterances). The recordings include reading

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.

	N	*T1*	*T2*	*T3*	*T4*	*T8*	*T9*	*T10*	*T11*	*S*
%	7.6	5.1	5.2	4.5	4.4	4.9	5.2	5.2	4.5	4.6
CL	69.1	78.1	80.4	83.2	82.1	80.0	70.7	80.4	79.0	69.7

Table 1. % marked words (strictness) for each rater; agreement with all other raters in % CL (unweighted average recall).

errors, repetitions of words, word fragments, and nonverbals. The total size of the vocabulary is 942 words. In this paper we concentrate on 28 OHM-pupils (15 male, 13 female) with age 10 – 11 who had been learning English for half a year only. For this subset of 1300 utterances (72 min. of speech, 8088 words), ratings are available from 10 experts; they labeled all words as correctly pronounced (default) or mispronounced. A German university student of English (graduate level, rater *S*) marked all pronunciation deviations. The instruction for 8 German teachers of English (raters *T1* – *T4*, *T8* – *T11*) and a British native teacher of English (rater *N*) was to mark those words, where the teachers would have stopped and corrected the student in class. *T2*, *T3*, and *T4* have labeled the data half a year later again, to measure some intra-rater agreement. Ratings of 12 teachers on the text level have been reported in [9].

Tab. 1 shows, that the strictness of all raters is similar; 4.4 – 5.2 % of the words were marked; only rater *N* marked 7.6 %. However, even teachers seem not to have a perfect agreement for the mispronounced words: Only *T8* and *T11* have an intersection of 267 out of 325 or 355 words. In this case the mean recall is 78 %, in all other cases it is lower. The agreement for correctly pronounced words, however, is high (recall 99 % for *T8* and *T11*). Consequently, only in the selection of wrongly pronounced words from the set of many non-native accentuated words teachers differ. If we build the union of mispronounced words from all raters we get 18 % of the words; if at least 5 should agree, we obtain 3.1 % only. A good compromise is the agreement of at least 3 raters: the 5.6 % of marked words are close to each rater's strictness. To measure the overall agreement of each teacher with all other teachers (mispronounced if marked by at least 3 teachers) we use the class-wise averaged recognition rate (CL) which is the unweighted mean of the recalls REC_w and REC_c for the two cases wrongly and correctly pronounced

$$\mathrm{CL} = 0.5(\mathrm{REC}_w + \mathrm{REC}_c). \qquad (1)$$

These values in Tab. 1, have a mean of 77.3 % CL, highest for *T3* and lowest for the student and the native annotator. Comparing pairs of raters (not: one vs. all others), agreement is on the average 70 % CL; teachers and their second ratings agree with 78 - 80 % CL.

The PF-STAR NATIVE-corpus (14.2 hours, vocabulary of 1740 words) contains British children recorded by the University of Birmingham [2]: 159 children, age 4 – 14. In the following the NATIVE data is used to train the acoustic models of the speech recognizer and some statistics necessary for pronunciation scoring.

3. FEATURES

To classify wrongly pronounced words, for each word a 176-dimensional highly redundant feature-vector is calculated. 113 features among them are prosodic features from our prosody module; a subset of these features has been introduced in [10]. Further 63 features have been developed additionally for pronunciation scoring; the pronfex module (**pron**unciation **f**eature **ex**traction) is described in [11]. For feature calculation a forced alignment of the sentences that should have been uttered (the reference sentences are known in our task) is required (Fig. 1, B) as well as phone based duration and energy statistics estimated from the NATIVE children's data.

#	Group	Best feature per group	CL
25	*ProsEne*	mean of the energy [1,2]	59.9
10	*ProsFFT*	energy FFT coefficient 0 [0,0]	58.4
26	*Prosf_0*	minimum of the f_0 [1,1]	53.6
22	*ProsPos*	position of f_0 onset [0,0]	**60.0**
7	*ProsDur*	normalized duration [-1,-1]	54.9
8	*ProsJit*	mean of jitter [1,1]	52.9
8	*ProsShim*	mean of shimmer [-1,-1]	53.0
7	*ProsPauses*	pauses after word [0,0]	58.5
2	*Pauses*	long pauses after word	53.5
3	*ROS*	rate-of-speech * duration of word	**62.2**
5	*DurLUT*	expected word duration	**60.9**
3	*DurScore*	prob. of observed duration / ROS	**61.7**
7	*Likelihood*	word-score (forced alignm.) / ROS	**64.3**
3	*LikeliRatio*	alignment vs. recog. word chain	57.4
3	*PhoneSeq*	bigram prob. of phones / ROS	59.1
4	*Accuracy*	phone correctness	58.7
2	*Confidence*	posterior score of word in reference	**60.1**
13	*PhoneConf*	maximum (cf. Eq. 2)	**63.0**
18	*Context*	context [-1,0] of *Likelihood*	**60.8**

Table 2. Feature groups of prosodic and pronunciation features. CL in % for the best single feature per group with the LDA classifier.

For the pronunciation features we further need the best recognized word chain from a speech recognizer that includes time alignment of the recognized words and phones (Fig. 1, A), since some features are based on a comparison of the reference and the recognized word or phone sequences. The speech recognizer is trained on native data. Whereas we could show in [11] that pronunciation features basically extract information independent of the learner's mother tongue, we now extend the word recognizer's vocabulary with 2533 words that contain numerous possible mispronunciations, which also include some typical German mistakes. This way, we can do without any additional phone recognizer in our online system. The pronunciation variations are based on a set of 140 phone confusions, which draws a distinction to carefully designed approaches to evaluate a single sentence with few selected mispronunciation models as used in Fig. 1, C. The evaluation on non-natives shows only 45.8 % word accuracy with 4-gram language modeling after remapping pronunciation variations to the original 942 words, since here the difficult tasks to recognize children's speech and to recognize non-native beginners of English are combined.

In addition, a phoneme bigram model has been estimated on the reference texts of the native and non-native data. Some further statistics employed for the feature calculation are more difficult to obtain: we compare on the phone level the forced alignment and the recognition result to get statistics of typical phone-confusions on correctly pronounced and on wrongly pronounced words, respectively. To estimate these confusion matrices, we have to use the NON-NATIVE data with the annotations described in Sec.2; however, we use leave-one-speaker-out (*loo*) feature calculation and classification, to ensure, that the test speaker and his phone confusions have never been observed during training. A third phone-confusion statistic has been estimated on native English data.

The Prosody Module. Prosodic features are based on the energy, the fundamental frequency (f_0), jitter, shimmer, duration and pauses. The features are calculated for each word, and, additionally, for some of the neighboring words to encode information from the context, e.g. [-2,-1] means, that the feature value is calculated from the two preceeding words, [0,0] indicates the current word, and [1,1] the succeeding word. 25 features are based on the energy (*ProsEne*) of the signal, e.g. maximum, minimum, mean, the regression and some normalized energy values. *ProsFFT* are the first 10 Fourier-coefficients of the energy trajectory within the respective word (approx. 1 - 8 Hz). Further 26 features are calculated from the fundamental frequency (*Pros*f_0): maximum, minimum, mean, the value at the onset/offset (beginning/end of voiced region), and the regression. The position (*ProsPos*) of e.g. the extrema of energy and f_0 encodes duration characteristics of a word (maximum, minimum, onset, etc., 22 features). Further 7 duration features (*ProsDur*) are calculated using the duration that is normalized by the rate of speech. 16 features are based on jitter and shimmer (*ProsJit*, *ProsShim*). Pauses before and after the respective word are described in 7 *ProsPauses*-features. Tab. 2 shows all feature groups.

The Pronfex Module. Further 63 features are provided for pronunciation scoring. Most are calculated for the current word (by default context [0;0]); the fluctuation is modeled with 18 selected *Context* features. The number of features per group is given in Tab. 2. Pronfex provides 2 further *Pauses*-features for long pauses. Rate-of-speech features (*ROS*) represent the number of phones per time, normalized in different ways. All other pronunciation features also differ per group often in the way of normalization. A detailed description can be found in [11]. *DurLUT* compares the observed and the expected duration based on duration statistics. *DurScore* gives the probability of the observed duration. The log-likelihood of the words in the reference is used in 9 *Likelihood* features; *LikeliRatio* compares the scores of the recognized word and the word in the reference. The probability of the recognized phone sequence given a phone bigram model is calculated in *PhoneSeq*. *Accuracy* is the phone accuracy or the phone correctness if insertions are not penalized. Further, from the recognizer *confidence* scores are obtained, e.g. the posterior probability of the word in the reference. Finally, the phone confusion features (*PhoneConf*) compare recognition and forced alignment on the phone level. It is analyzed, whether the observed phone confusion is better represented by the confusion matrix of wrongly pronounced words ($\mathbf{M}_1$) or the confusion matrix of correctly pronounced words ($\mathbf{M}_0$). The latter matrix is for some features replaced with the confusion matrix of phones uttered by natives ($\mathbf{M}_E$). As word features, the maximum, minimum, mean, etc. of the phone based observations are used, e.g.

$$\textit{PhoneConf-mean}: \quad \frac{1}{N}\sum_j P(q_j|p_j,\mathbf{M}_1)/P(q_j|p_j,\mathbf{M}_0) \quad (2)$$

where q_j is the recognized phone and p_j the phone in the reference ($j = 1 \dots N$: indices of phones in the word).

4. ADABOOST

In this paper it is investigated which subset of the 176 correlated features is most important. A widely used algorithm is ADABOOST, that selects those weak classifiers that use complementary information and combines them to a strong classifier. ADABOOST has been introduced by Freund and Schapire [12] and has turned out to be very robust against overfitting to the training data [13]. In our case each weak classifier is trained on exactly one of our 176 features. The classifier consists of a threshold and a sign to determine the class, either wrongly or correctly pronounced. To detect mispronunciation, we select different numbers of features (weak classifiers) that are finally combined to a strong classifier and evaluated on the test data set. A weak classifier $h_t(\mathbf{x}_i)$ for the word $\mathbf{x}_i$ returns 1, if mispronunciation is classified and 0 else. In the first step, we calculate optimal thresholds for each weak classifier on the training data; the criterion is CL (Eq. 1). The algorithm is the following, starting with $t = 0$:

$\bar{\alpha}$	Group	Selected feature
0.82	*PhoneConf*	mean (defined in Eq. 2)
0.52	*Likelihood*	word-score (forced alignment)
0.28	*Context*	context [-1,0] of *Confidence*
0.27	*ProsFFT*	FFT coefficient 1 [0,0]
0.26	*DurLUT*	scatter of phone duration deviation
0.24	*Confidence*	posterior score of word in reference
0.22	*ProsFFT*	FFT coefficient 0 [0,0]
0.21	*Pros*f_0	regression of the f_0 [-1,0]
0.21	*PhoneSeq*	bigram prob. of phones / #phones
0.19	*ProsPos*	position of the maximal f_0 [1,1]
0.17	*ProsEne*	minimum of the energy [-2,-1]
0.17	*ProsDur*	normalized duration [-1,0]
0.16	*PhoneConf*	maximum confusion score (cf. Eq. 2)
0.15	*ProsEne*	mean of the energy [1,2]
0.15	*PhoneConf*	minimum confusion score (cf. Eq. 2)

Table 3. Top 15 features selected with ADABOOST and ranked with their mean α-values after Eq. 3.

1. A weight $w_{0,i}$ is assigned to each word i of the training data, so that the weights of either class sum up to 0.5.
2. Choose the weak classifier $h_t(.)$ with lowest error ϵ_t: Words that are wrongly classified contribute with $w_{t,i}$ to the error.
3. Use greater weights for all wrongly classified words:

$$w_{t+1,i} = w_{t,i}\frac{1-\epsilon_t}{\epsilon_t} \quad ; \quad \alpha_t = \log(\frac{1-\epsilon_t}{\epsilon_t}) \qquad (3)$$

4. Normalize the weights; $t = t + 1$; goto 2.

Due to the new weights, the second best feature uses complementary information and so on. In the end, the strong classifier is obtained by a linear combination of all selected weak classifiers:

$$\sum_t \alpha_t h_t(\mathbf{x}) \quad \geq \quad \frac{1}{2}\sum_t \alpha_t \qquad (4)$$

shows that $\mathbf{x}$ is mispronounced.

5. EXPERIMENTS AND RESULTS

All experiments are conducted on the non-native children data using leave-one-speaker-out (*loo*) evaluation on 28 speakers. Words are labeled as mispronounced if at least 3 raters agree. First, single features are evaluated with the LDA classifier (Tab. 2). Best results are obtained with pronunciation features, in particular with the log-likelihood of the recognized word, normalized by the rate-of-speech, and the maximum phone confusion per word *PhoneConf-max* (cf. Eq. 2, maximum instead of mean). Prosodic features have low classification rate, however, they will be a useful extension to the pronunciation features.

Feature selection with ADABOOST is also performed in *loo* mode. In each *loo* iteration *PhoneConf-mean* is the optimal feature (Eq. 2) and the second best is always *Likelihood* except in one case, where *DurLUT* wins. The three features that follow next contain in most cases *ProsEne*, *ProsFFT* and a *Context* feature based on the confidence (posterior probability of the reference word in the recognizer's wordgraph).

Now, the ADABOOST results of all *loo* iterations are merged using the scores α_t from Eq. 3. For each feature the mean of the respective values α_t is built from all 28 *loo* iterations. Those $\bar{\alpha}_t$ are used to re-sort the joint feature lists. Tab. 3 shows the top 15 of the new

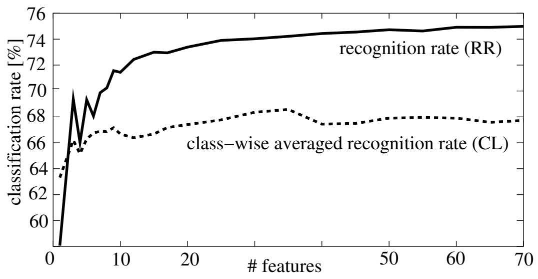

Fig. 2. Classification rates on the test data for different numbers of features selected with ADABOOST (*loo*-evaluation).

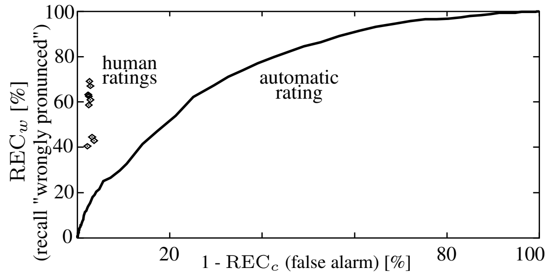

Fig. 3. ROC evaluation based on 35 features.

ranking. *PhoneConf* appears even three times and many prosodic features are selected. Not visible in the table are *ProsPauses* (top 19, $\bar{\alpha} = 0.12$), *ProsShim* (top 29, $\bar{\alpha} = 0.08$), *ROS* (top 33, $\bar{\alpha} = 0.07$), *Accuracy* (top 36, $\bar{\alpha} = 0.06$), and *ProsJit* (top 45, $\bar{\alpha} = 0.05$). Some feature groups are not selected at all in the 70 best features' list and possibly replaced with features with similar classification performance. Using 15 features, the classification rate is 66.7 % CL, with 35 features 68.6 % CL. The overall recognition rate RR (% correctly classified) rises from 73.0 to 74.2 %. Fig. 2 shows the performance dependent on the number of features. Even with large feature numbers no overfitting to the training data is observed: classification rate rises on the test set. ROC-evaluation is shown in Fig 3.

Comparing the *loo* iterations we found, that similar features are selected. In [11], it could be shown that classification with pronunciation features even works, if we train and test on speakers with different native tongues. In the following we investigate the dependency on individual raters. It turned out, that for each rater similar features are selected, in particular *PhoneConf* and *Likelihood* are always among the top features. The only exception is rater *S*: One of the *ProsPauses*-features is the best feature, followed by a *PhoneConf*-feature that is calculated on the native phone-confusion matrix (M_E) instead of using the correctly pronounced non-native words (M_0). The rating is not like a teacher would do, but a precise comparison with native speech that also considers pauses and hesitations as uncertainty of the L2-learner.

6. DISCUSSION

With the ADABOOST algorithm a subset of the 176 prosodic and pronunciation features was selected. The 15 best features combine uncorrelated information, including prosody and information from a speech recognizer. They consider phone confusions that differ in automatic speech recognition for correct and wrong pronunciation. Confidence measures are used from the speech recognizer and log-likelihood scores from the forced alignment. The energy of the word is analyzed with the lowest Fourier-coefficients and mean/minimum values from the preceeding and succeeding words. The prior probability of the observed phone sequence (it differs due to the 2533 mispronunciation models of the speech recognizer) is estimated with bigram statistics. As for the f_0, slope and position of the maximum are analyzed. The normalized duration and the deviation from average durations estimated on native speakers is considered. With 35 features 68.6 % CL (74.2 % RR) was achieved. This is 89 % of the agreement of human experts (77.3 % CL). However, with comparable false alarm rate only 27 % of the average teachers' recall for mispronunciation is reached (47 % of *S*). Teachers have a high agreement on correctly pronounced words, but a surprisingly low hit-rate on mispronounced words, even if we compare teachers with themselves. Currently, this feature set is being integrated in our *Caller* system. It is combined with a common approach that is based on acoustic models with selected, typical German mispronunciations.

7. REFERENCES

[1] D. Herron, W. Metzel, E. Atwell, R. Bisiani, F. Daneluzzi, R. Morton, and J. Schmidt, "Automatic localization and diagnosis of pronunciation errors for second-language learners of English," in *Proc. Eurospeech*, Budapest, 1999, pp. 855–858.

[2] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, " The PF-STAR Children's Speech Corpus ," in *Proc. Eurospeech*, 2005, pp. 2761–2764.

[3] A. Heßler, "Eine Client-Server Anbindung zur automatischen Aussprachebewertung für das Projekt "Caller"," Diploma Thesis, University of Erlangen-Nuremberg, Germany , 2006.

[4] M. Eskenazi, Y. Ke, J. Albornoz, and K. Probst, "The Fluency Pronunciation Trainer: Update and User Issues," in *Proc. of InSTIL*, Dundee, 2000.

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

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

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

[8] 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.

[9] C. Hacker, A. Batliner, S. Steidl, E. Nöth, H. Niemann, and T. Cincarek, " Assessment of Non-Native Children's Pronunciation: Human Marking and Automatic Scoring ," in *Proc. SPEECOM*, 2005, vol. 1, pp. 123 – 126.

[10] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to Find Trouble in Communication," *Speech Comm.*, vol. 40, pp. 117–143, 2003.

[11] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann, " Pronunciation Feature Extraction ," in *Pattern Recognition, 27th DAGM Symposium* , 2005, pp. 141–148.

[12] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proc. EuroCOLT*, London, UK, 1995, pp. 23–37.

[13] Y. Freund and R. Schapire, "A short introduction to boosting," *J. of Japanese Soc. for Artif. Intel.*, vol. 14, pp. 771–780, 1999.