Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 23 (2009) 65–88
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Automatic pronunciation scoring of words and sentences independent from the non-native's first language

Tobias Cincarek [a,*], Rainer Gruhn [a,1], Christian Hacker [b], Elmar Nöth [b], Satoshi Nakamura [a]

[a] *ATR Spoken Language Translation Research Labs, 2-2-2 Hikaridai, Keihanna Science City, 619-0288 Japan*
[b] *Institute for Pattern Recognition, Friedrich-Alexander University Erlangen-Nuremberg, Germany*

Received 6 October 2006; received in revised form 8 January 2008; accepted 5 March 2008
Available online 12 March 2008

Abstract

This paper describes an approach for automatic scoring of pronunciation quality for non-native speech. It is applicable regardless of the foreign language student's mother tongue. Sentences and words are considered as scoring units. Additionally, mispronunciation and phoneme confusion statistics for the target language phoneme set are derived from human annotations and word level scoring results using a Markov chain model of mispronunciation detection. The proposed methods can be employed for building a part of the scoring module of a system for computer assisted pronunciation training (CAPT). Methods from pattern and speech recognition are applied to develop appropriate feature sets for sentence and word level scoring. Besides features well-known from and approved in previous research, e.g. phoneme accuracy, posterior score, duration score and recognition accuracy, new features such as high-level phoneme confidence measures are identified. The proposed method is evaluated with native English speech, non-native English speech from German, French, Japanese, Indonesian and Chinese adults and non-native speech from German school children. The speech data are annotated with tags for mispronounced words and sentence level ratings by native English teachers. Experimental results show, that the reliability of automatic sentence level scoring by the system is almost as high as the average human evaluator. Furthermore, a good performance for detecting mispronounced words is achieved. In a validation experiment, it could also be verified, that the system gives the highest pronunciation quality scores to 90% of native speakers' utterances. Automatic error diagnosis based on a automatically derived phoneme mispronunciation statistic showed reasonable results for five non-native speaker groups. The statistics can be exploited in order to provide the non-native feedback on mispronounced phonemes.

Keywords: Non-native speech; Pronunciation assessment; Sentence scoring; Word scoring; Mispronunciation detection; Phoneme mispronunciation statistic

* Corresponding author. Present address: Yahoo Japan Research, Yahoo Japan Corporation, Midtown Tower, 9-7-1 Akasaka, Minato-ku, 107-6211 Japan.
E-mail address: tcincare@yahoo-corp.jp (T. Cincarek).
[1] Presently, the author is with Harman/Becker and University of Ulm, Germany.

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.03.001

1. Introduction

Systems for computer assisted language learning (CALL) are intended to provide learners of a second language a medium to improve their skill in a foreign language without the presence of a human teacher. While self-study in grammar or vocabulary can be successful, feedback is especially important for pronunciation training (Neri et al., 2002b). From a pedagogical point of view, a system for computer assisted pronunciation training (CAPT) should provide the student an overall assessment of pronunciation quality to verify correctness, pinpoint certain rather than highlight all mistakes, and possibly suggest a remedy (Neri et al., 2002a). In order to provide feedback without human assistance, methods for automatically scoring the pronunciation quality at different levels of granularity are required. This paper presents a pronunciation scoring method applicable independently of the non-native's first language. The focus is on technological aspects of scoring the pronunciation quality of words and sentences. Furthermore, a method to derive feedback about mispronunciations at the phoneme level from word level scoring is proposed and explained using examples from five non-native speaker groups.

Research on pronunciation scoring has been carried out for the phoneme (Witt and Young, 2000), sentence (Neumeyer et al., 2000; Franco et al., 2000; Teixeira et al., 2000) and speaker level (Bernstein et al., 1990; Minematsu, 2004). Prevalent are approaches based on speech recognition technology. Features describing the pronunciation quality are extracted from the output of a speech recognizer, in particular forced-alignment and recognition result. Examples for features are the posterior probability of phonemes, phoneme model likelihood, duration score, rate of speech, recognition accuracy and duration of pauses. To extract these so-called pronunciation features, the reference transcription, an acoustic model and duration statistic for each phoneme of the target language, and possibly a language model for recognition are required.

To validate an automatic scoring result, a human reference is required. Such a reference can be obtained by a human evaluation of non-native speech material. Examples for an evaluation on phoneme and sentence level can be found in literature: sentences are labeled on a discrete scale (e.g. from 1 to 5) indicating an utterance's overall pronunciation quality (Cucchiarini et al., 2000); phonemes are classified either as correctly pronounced or as mispronounced (Witt and Young, 2000).

An approach for sentence (and speaker) level scoring in case the spoken text is unknown is proposed in Moustroufas and Digalakis (2007). It requires an additional acoustic and (phoneme) language model for each possible first language of the non-native.

Speaker-level scoring may be appropriate to determine the overall pronunciation skill of a non-native speaker, e.g. in language testing applications. Scoring at the sentence level provides an immediate assessment of overall utterance correctness. By averaging the scores of utterances over fixed time intervals, the foreign language student's overall progress can be measured. However, in order to provide the learner a more detailed feedback on certain mistakes in pronunciation, scoring should also be carried out at the word or phoneme level. Although immediate feedback at the phoneme level is important, there are several reasons to consider also direct pronunciation assessments at the word level. The reliability of human ratings and scores decreases the finer the level of granularity, since shorter speech segments contain less information. Furthermore, mispronunciations may not only arise from the incapability of the non-native to articulate speech sounds which are not part of his native language, but they may also rely on phonetic contexts not occurring in the speaker's native language and mistakes in transferring a given grapheme sequence into the correct phoneme sequence due to phonotactical rules which are not used in the target language.

In this paper, an approach for pronunciation scoring independent of the learner's first language is proposed and evaluated. Although the target language is English, the proposed methods can easily be applied to different target languages as long as reference speech data are available. The main investigation target is the detection of mispronounced words, since only little work has been done on scoring the pronunciation of words directly. A set of word level features is developed by applying sentence level features, applying confidence measures and combining phoneme level features.

Furthermore, a data-driven approach for automatic error diagnosis is investigated. In comparison to knowledge-based approaches (Herron et al., 1999; Tsubota et al., 2002; Park and Rhee, 2004; Ito et al., 2005) it has the advantage of being in principle independent from the non-native's first language. It is shown that detailed feedback about mispronounced phonemes can be derived using the word level scoring result.

Using a Markov chain model of mispronunciation detection, the mispronunciation probability of phonemes are estimated from a statistic of mispronounced words. Findings from automatically derived statistics about phoneme mispronunciations and phoneme confusions are reasonable for five groups of non-natives against the background of phonological comparisons between the non-native's first language and the target language.

Apart from word scoring, sentence scoring is also considered. Besides investigating an improved feature set, a more meaningful evaluation measure is proposed. The problem when only using the correlation coefficient to determine the parameters of a scoring function e.g. by using linear regression is that the score range and actual score values may deviate much from the desired target.

The outline of this paper is as follows: Section 2 describes two non-native speech databases: multi-accented English adult speech and German-accented English children speech. The annotations of the databases are analyzed for various aspects, e.g. the inter-rater reliability at word, sentence and speaker level in Section 3. Sentence level scoring is considered in Section 4, word level scoring in Section 5. Experimental conditions are explained in Section 6. Experimental results for both databases and word and sentence level scoring are given in Section 7. In Section 8 automatic error diagnosis based on a word mispronunciation model is investigated. From a statistic of mispronounced words, it is possible to derive feedback about mispronounced phonemes for the foreign language student. Conclusions are drawn in Section 9.

2. Data and labels

The development of a pronunciation scoring module requires non-native speech with annotations regarding pronunciation quality. In the following, two non-native speech databases are described: multi-accented non-native English speech from German, French, Japanese, Indonesian and Chinese adults (ATR SLT data), and non-native English speech from German school children (PF-STAR data). All speech data are annotated with tags for mispronounced tokens at the word level and discrete ratings for overall pronunciation quality at the sentence level.

If there are two or more reference labels for the same item, they have to be combined in a meaningful way. For the sentence level this is achieved by averaging the ratings of all human evaluators. A word is considered mispronounced if it was marked by two or more of the evaluators. Otherwise, the word's pronunciation is assumed to be correct.

2.1. ATR SLT non-native database

At the spoken language translation (SLT) research laboratories of ATR international a non-native English speech database was collected (Gruhn et al., 2004). It contains foreign accented English speech from 96 non-native English speakers. The first language of most speakers is German, French, Indonesian, Chinese and Japanese. Each subject had to read the 48 phonetically rich sentences of the TIMIT (Garofolo et al., 1993) SX set, credit card numbers and hotel reservation dialogs. Each subject was able to listen to recorded utterances. The recording of a sentence was repeated, either if the subject was not satisfied or if the subject misread the reference sentence, but only in case of insertions or deletions of whole words. This measure assures a one-to-one correspondence between the word sequence of the reference sentence and a student's spoken word sequence for proper classifier training and valid word and sentence level pronunciation scoring. Otherwise the spoken word sequence would have to be recognized automatically which is difficult especially for non-native speech.

The phonetically rich sentences were employed for investigation of a pronunciation scoring algorithm, since they cover all phonemes and many phoneme contexts in order to assure enough phonemic diversity. There are 4608 utterances (96 speakers times 48 sentences) corresponding to 6.4 h of non-native speech in total. The data were divided into four speaker-disjoint subsets containing 1152 sentences each. Each subset was evaluated by three or four native speakers with professional English teaching experience from the US and Canada. The pronunciation assessments at the sentence level are on a discrete scale from '1' for native-like pronunciation to '5' for unintelligible pronunciation. Furthermore, words perceived as mispronounced were marked (binary label). Table 1 shows the frequency of each reference label on word and sentence level. Most utterances are labeled '2' and '3', indicating most speakers have an intermediate pronunciation skill. The relative share of words

Table 1
Distribution of the reference labels at the word and the sentence level (ATR data)

	Sentence					Word	
Label	1	2	3	4	5	Correct	Mispronunciation
Frequency	382	1791	1744	623	68	34392	3528
	8.3%	38.9%	37.8%	13.5%	1.5%	90.7%	9.3%

Table 2
Distribution of the reference labels at the word and the sentence level (PF-STAR data)

	Sentence					Word	
Label	1	2	3	4	5	Correct	Mispronunciation
Frequency	473	1410	458	117	61	9671	391
	18.8%	56.0%	18.2%	4.6%	2.4%	96.1%	3.9%

considered mispronounced is 10% forming a solid basis to build a scoring system which should be able to pinpoint mistakes.

Evaluators were instructed to consider segmental aspects, strong non-native accent, long between-word pauses in their rating decision but to ignore sentence intonation. Before starting proper evaluation, each evaluator had to listen to a uniform set of 22 utterances from 22 non-natives. The utterances were selected in order to have at least one representative for each foreign accent and one speaker with low, medium and high speech recognition accuracy. This measure was intended to give the evaluators a clue for the upper and lower bound of the rating scale.

For validation purposes, speech of seven native speakers is also available. All words in the natives' utterances are assumed to be pronounced perfectly. The data are employed to validate the automatic scoring algorithm. It is expected, that the native speakers obtain scores indicating high pronunciation quality.

2.2. PF-STAR non-native database

At the Institute of Pattern Recognition, University Erlangen-Nuremberg, Germany the PF-STAR (Batliner et al., 2005) non-native database was collected. It contains non-native English speech from 57 German children (10–15 years of age) of two high-schools. Most of the subjects have been learning English for about six months, all others for about eighteen months. The children read sentences from their English text book, some phrases and isolated words. The database contains 4627 utterances (3.4 h of speech) in total.

2519 utterances of the database with annotations at the word and the sentence level are employed for experiments. The PF-STAR database was collected independently from the ATR data. Word and sentence level annotations are of the same kind as in the ATR database. The speech data were annotated by eight human evaluators, among them three experienced and three trainee teachers of English, who are native Germans. The evaluators were instructed to mark those mispronounced words, which they would like to correct in the first instance.

Table 2 shows the frequency of each reference label. Most utterances are rated '2' or '3' being similar to ATR data. The percentage of words considered mispronounced was approx. 4%, which is less than half in comparison to ATR data.

3. Analysis of the human evaluation

For evaluating the inter-rater reliability of the pronunciation labels at word, sentence and speaker level, the correlation coefficient $C(X,Y)$ is employed. It is defined as

$$C(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_X)^2\sum_{i=1}^{n}(y_i-\mu_Y)^2}} \tag{1}$$

where the values x_i, y_i of the random variables X, Y are corresponding pronunciation annotations for the same item (e.g. sentence) with index i assigned by two human raters X and Y. μ_X denotes the mean of random variable X. To account for the ratings of three or more human evaluators, the open correlation for each human evaluator X_j is used additionally. It is defined as the correlation between one random variable and the average of the remaining variables if there are $k \geqslant 3$ human evaluators in total

$$C_{\text{open}}(X_j) = C\left(X_j, \frac{1}{k-1}\sum_{i \neq j} X_i\right) \tag{2}$$

3.1. ATR data

Table 3 shows the results of the human evaluation for both reliability measures at each level of granularity. Speaker level ratings were obtained by averaging all sentence ratings available for a speaker. It is obvious, that the higher the level of assessment, the higher the reliability. This is natural, since the rating decision of an evaluator is based on more speech material.

There have been a few evaluator pairs with a low correlation at the word and sentence level. This is due to the subjectiveness of each evaluator. The strictness of an evaluator is a characteristic element of subjectiveness. For example, in case of the evaluator pair with the lowest correlation (0.16) at the word level, one evaluator only marked 2%, but the other as much as 28% of the words. For the pair with highest correlation (0.52) the corresponding evaluators marked 12% and 14% of the words, respectively. The same applies to the sentence level, since there is a relationship between the average number of marked words and the sentence level rating (Table 4). The correlation between the number of marked words in a sentence and the corresponding sentence rating was 0.63 on average.

Additional statistics can be derived from the pronunciation labels. As Fig. 1 shows, the higher the number of phonemes in a word, the higher the relative marking frequency. This is easy to understand, since an evaluator may mark a word if there is at least one mispronounced phoneme and the possibility for mispronunciation increases with the number of phonemes. The words ‘extra’, ‘exposure’, ‘exam’ and ‘box’ in every speaker’s material were considered as mispronounced by every evaluator. Other mispronounced words, which had a relative marking frequency greater than 0.75, were ‘mirage’, ‘centrifuge’, ‘bugle’, ‘frantically’, ‘oasis’ and ‘purchase’.

Table 3
Minimum, average and maximum of pair-wise inter-rater correlation and the open inter-rater open correlation on word, sentence and speaker level (ATR data)

Level	Correlation			Open correlation		
	Minimum	Average	Maximum	Minimum	Average	Maximum
Word	0.16	0.34	0.52	0.27	0.44	0.57
Sentence	0.28	0.49	0.65	0.44	0.60	0.70
Speaker	0.85	0.91	0.97	0.88	0.94	0.98

Table 4
Relationship between the sentence rating and the number of marked words in the sentence

# Marked words	0	1	2	3	4	$\geqslant 5$
Rating 1	0.96	0.04	0.00	0.00	0.00	0.00
Rating 2	0.70	0.28	0.02	0.00	0.00	0.00
Rating 3	0.30	0.44	0.21	0.04	0.01	0.00
Rating 4	0.09	0.27	0.37	0.19	0.07	0.01
Rating 5	0.00	0.06	0.17	0.33	0.35	0.10

Each value means the relative frequency of sentences with a certain rating and a certain number of words marked (ATR data).

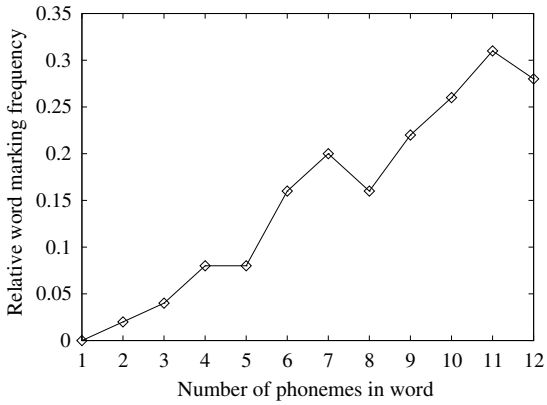

Fig. 1. Relationship between the number of phonemes in a word and its relative marking frequency (ATR data).

Table 5
Pair-wise inter-rater correlation and open inter-rater correlation on word level (PF-STAR data)

Level	Correlation			Open correlation		
	Minimum	Average	Maximum	Minimum	Average	Maximum
Word	0.28	0.41	0.56	0.47	0.59	0.67

3.2. PF-STAR data

Mispronounced words were marked by eight human evaluators. There is a correlation of 0.5 between the number of marked words and the sentence level ratings. The minimum, maximum and average correlation between word level annotations of each rater pair and the open correlation are shown in Table 5. The average inter-rater correlation is higher than for the ATR data. One reason is, that the strictness among the evaluators of the PF-STAR data was almost uniform (about 4% marked). Another reason might be the fact that the PF-STAR data are from a homogenous speaker group (German school children) and most of the annotators know each other and presumably some of the non-native subjects (English teachers from the same school as the non-native children). The ATR database contains multi-accented non-native speech data from subjects a priori unknown to the evaluators.

Three teachers evaluated the same material twice. Their intra-rater correlation is 0.56, 0.59 and 0.63, respectively, higher than the average inter-rater correlation.

Examples of words with a mispronunciation probability equal or greater than 0.75, are ‘hotel’, ‘intelligent’, ‘casual’, ‘example’, ‘fortunately’, ‘comprehensive’, ‘transparent’, ‘garage’ and ‘confusion’.

4. Sentence scoring

Besides employing sentence level features from literature, several features are modified and some new features are proposed. The feature extraction is based on the forced-alignment and the phoneme recognition output of the target utterance with an acoustic model trained on native speech. This native speech serves as reference material for the desired quality and characteristics of pronunciation.

4.1. Sentence level pronunciation features

In Table 6 variables and symbols used in feature definitions are summarized. Feature extraction is carried out separately for each sentence $\vec{S}$. A sentence can be said to consist of either N phoneme segments or M word segments, which are also made up of phoneme segments. It is assumed, that there are no intra-word pauses,

Table 6
Definition of variables and symbols for sentence level pronunciation features

Entity	Symbol	Definition
Sentence	$\vec{S}$	Word sequence $(W_1,\ldots,W_M)$ Phoneme sequence $(p_1,\ldots,p_N)$ Phoneme segments $(\vec{X}_1,\ldots,\vec{X}_N)$
Segment	$\vec{X}$	Frame sequence $(\vec{x}_1,\ldots,\vec{x}_T)$
Frame	$\vec{x}$	Acoustic features $(x_1,\ldots,x_d)$
Duration	T	Phoneme segment duration $(\vec{X})$
	D	Word segment duration (W)
	T_S	Total sentence duration
# Phonemes	N	Num. of phoneme segments in $\vec{S}$
# Words	M	Num. of word segments in $\vec{S}$
Speaking rate	$R^{(ph)}$	# Phonemes (N)/time (T_S)
	$R^{(wd)}$	# Words (M)/time (T_S)

but only inter-word pauses. The duration of a phoneme segment $\vec{X}_i$ with label p_i is denoted as T_i. Word durations are denoted as D_j. The total duration T_S of a sentence is defined as the duration of all phonemes plus inter-word pauses in the sentence. Leading and trailing silence segments are ignored. The rate of speech (ROS) is a measure of the speaking rate. It can be defined as the number of phonemes, syllables or words per time.

The rate of speech can be used as pronunciation feature. However, experiments revealed that there is a higher correlation for the reciprocal phoneme-based rate of speech, i.e. the mean phoneme duration

$$(\text{MeanPhDur})\ \mathscr{R} = \frac{1}{R^{(ph)}} \tag{3}$$

Another feature is the duration score (Neumeyer et al., 2000) to measure deviations from the duration characteristics typical for native speech. The score is calculated as the sum of the sentence's phoneme models' duration log-likelihood

$$(\text{DurScore})\ \mathscr{D} = \frac{1}{N}\sum_{i=1}^{N} \log P_{dur}^{(ph)}(T_i * R^{(ph)}|p_i) \tag{4}$$

A phoneme duration probability density function (pdf) can be estimated from transcribed native speech data. Instead of approximating the pdf with a histogram, the log-normal density function

$$P_{dur}^{(ph)}(t|p) = \frac{1}{t\sqrt{2\pi\sigma_p^2}} \exp\left[-\frac{(\log t - v_p)^2}{2\sigma_p^2}\right] \tag{5}$$

is employed, since phoneme durations are distributed log-normal. The parameters v_p and σ_p are obtained by maximum-likelihood estimation based on ROS-normalized duration samples for each phoneme. This normalization is necessary in order to account for variations of the speaking rate.

The acoustic model likelihood $L(\vec{X}) = \log P(\vec{X}|\lambda_p)$ can be considered as a measure of acoustic similarity between the target speech and the context-independent acoustic model λ_p for phoneme p. Here, the original definition of the likelihood-based pronunciation feature (Neumeyer et al., 2000) is modified by additionally normalizing with the rate of speech, since the correlation to human ratings increased further

$$(\text{SentLh1})\ \mathscr{L} = \frac{1}{N}\sum_{i=1}^{N} \frac{L(\vec{X}_i)}{T_i * R^{(ph)}} \tag{6}$$

To calculate feature $\mathscr{L}$, each segment's likelihood is divided by its actual duration. Alternatively, normalization is possible by dividing with the expected (phoneme or word) duration. This is realized for the following new pronunciation feature:

$$(\text{SentLh2})\ \mathscr{E} = \frac{1}{M}\sum_{j=1}^{M}\frac{L(W_j)}{D_j^{(e)} * R^{(\text{wd})}} \tag{7}$$

$L(W_j)$ denotes the sum of phoneme model log-likelihoods of word W_j. An estimate for the expected word duration $D_j^{(e)}$ is the sum of the mean duration of the phonemes of word W_j.

Besides the phoneme likelihood, the phoneme posterior probability $P(p_i|\vec{X})$ is a promising pronunciation feature. In (Neumeyer et al., 2000) it was shown to be the feature with highest correlation to human ratings. Its calculation was simplified to the likelihood ratio

$$L_r(X_i|p_i) = \sum_{t=1}^{T_i}\log\frac{P(\vec{x}_t|p_i)}{P(\vec{x}_t|q_t^*)} \tag{8}$$

where q_t^* is the name of the model with highest likelihood given frame $\vec{x}_t$, i.e. $q_t^* = \text{argmax}_{q\in Q}P(\vec{x}_t|q)$. Q is the phoneme set of the target language. In practice, q_t was obtained by unconstrained phoneme recognition. Thus a likelihood ratio score was obtained for each phoneme segment. These scores are normalized by the actual segment duration, summed up and finally divided by the number of segments N. Here, the feature is modified to

$$(\text{LhRatio})\ \mathscr{K} = \frac{\sum_{i=1}^{N}L_r(\vec{X}_i)}{\sum_{i=1}^{N}T_i^{(e)} * R^{(\text{ph})}} \tag{9}$$

i.e. normalizing the segments posterior scores by the product of the speaking rate and the expected segment duration $T^{(e)}$, since the correlation to human ratings increased further.

An indicator of how good an utterance can be recognized is the phoneme or word accuracy. The former is a better measure, since it is based on a larger number of tokens. The accuracy can be calculated as the normalized minimum-edit-distance

$$(\text{PhAcc})\ \mathscr{A} = \frac{\text{MinEditDist}(\vec{q},\vec{p})}{\max\{|\vec{q}|,|\vec{p}|\}} \tag{10}$$

The distances of insertions, deletions and substitutions are uniformly set to one. $|\vec{p}|$ means the number of phonemes in the phoneme reference vector $\vec{p}\cdot\vec{q}$ denotes the phoneme recognition hypothesis. $\mathscr{A}$ is zero if reference and hypothesis are identical and greater than zero, if there are recognition errors.

Being unsure about a word's pronunciation may introduce inter-word pauses. Consequently, it is worth considering the total duration (PauseDur) $\mathscr{P}$ of inter-word pauses (Teixeira et al., 2000) within a sentence as a feature.

As a further new pronunciation feature the probability of the recognized phoneme sequence $\vec{q}$ given an n-gram language model (LM) is employed. The LM should be trained on canonic phoneme transcriptions of valid sentences of the target language, because a foreign language student should acquire standard pronunciation

$$(\text{PhSeqLh})\ \mathscr{M} = \frac{1}{R^{(\text{ph})}}\log P(\vec{q}|\text{LM}) \tag{11}$$

Each pronunciation feature is intended to measure certain aspects of pronunciation. $\mathscr{R}, \mathscr{D}$ and $\mathscr{P}$ are measures for temporal characteristics like the fluency of a speaker. $\mathscr{L}$ and $\mathscr{K}$ are intended to measure the segmental quality. $\mathscr{M}$ and $\mathscr{A}$ can be considered as indicators for both kinds of characteristics. Other prosodic features, e.g. based on fundamental frequency, had a rather low correlation with human ratings (Teixeira et al., 2000).

4.2. Scoring method

Fig. 2 shows the two approaches examined for sentence scoring. The Gaussian classifier with maximum likelihood decision rule itself can provide a discrete scoring result (hard scoring). A continuous scoring result (soft scoring) can be obtained by calculating the expected score value from the likelihood of the class models and the class prior probabilities. The latter are considered to be distributed uniformly. Another approach for

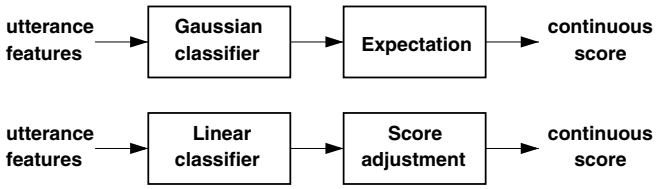

Fig. 2. Approaches to sentence scoring.

soft scoring is to use a linear combination of the pronunciation features. The weighting coefficients for each feature can be estimated using linear regression.

4.3. Score adjustment

In experiments it was observed that the result of soft scoring when using a linear feature combination is skewed. To improve scoring accuracy a linear

$$x = g(s, \vec{a}) = a_0 + a_1 s$$

and a multiplicative polynomial transformation

$$s' = x * f(x, \vec{b}) = b_0 x + b_1 x^2 + b_2 x^3 + \cdots + b_k x^{k+1}$$

can be applied in series to the raw scoring output s of the linear classifier to obtain a more accurate score s'.

The linear transformation is able to adjust the mean and the slope of the regression function, the second transformation is employed to correct certain non-linear distortions.

The parameters $\vec{a} = (a_0, a_1)$ of the linear transformation $g(s, \vec{a})$ can be obtained by linear regression based on the scoring result and reference labels for the training data. However, instead of estimating $\vec{a}$ to map a score s to its corresponding reference x, the variables s and x are interchanged, i.e. the parameters $\vec{a}' = (a'_0, a'_1)$ of $s = g(x, \vec{a}')$ are estimated. The coefficients $\vec{a}$ of the desired linear mapping can then taken from the reciprocal function of $g(x, \vec{a}')$

$$a_0 = -\frac{a'_0}{a'_1} \quad \text{and} \quad a_1 = \frac{1}{a'_1} \tag{12}$$

Let μ_h be the average score for a sentence with reference rating h. It was often the case that either $\mu_h > h$, i.e. the actual score values for tokens labeled h are too large, or $\mu_h < h$, i.e. the score values were too small. In order to bring them closer to their reference value, they have to be multiplied with a value either smaller or larger than 1. The multipliers for the whole score range can be defined by polynome $f(x, \vec{b})$.

The coefficients $\vec{b} = (b_0, b_1, \ldots, b_k)$ of f can be determined using interpolation with Newton's method. The desired multiplicative transformation is the polynome f fitting through the k points $\left(h, \frac{h}{\mu_h}\right)$.

5. Word classification

Mispronounced words could be detected using a continuous word score as in sentence scoring and a threshold to decide on mispronunciations. Since the purpose is in the end to discriminate correctly pronounced words (correct) from mispronounced words (wrong), the issue is considered as a two-class classification problem in the following.

5.1. Word level pronunciation features

Any feature defined for the sentence level can be applied to the word level in principle, since sentences consisting of only one word are valid. However, preliminary investigations revealed, that features with high quality for the sentence level are not necessarily good for the word level. Table 7 briefly explains variables and symbols employed for feature definitions.

Table 7
Definition of variables and symbols for word level pronunciation features

Entity	Symbol	Definition
Word sequence	$\vec{W}$	Word labels $(W_1,\ldots,W_M)$
	$\vec{O}$	Acoustic observation $(O_1,\ldots,O_M)$
Word	W	Phoneme labels $(p_1,\ldots,p_n)$
	O	Acoustic segments $(\vec{X}_1,\ldots,\vec{X}_n)$
Phoneme segment	$\vec{X}$	Frame sequence $(\vec{x}_1,\ldots,\vec{x}_T)$
		Reference labels $(p_1,\ldots,p_T)$
		Hypothesis labels $(q_1^*,\ldots,q_T^*)$
# Phonemes	n	Number of phonemes in word W

Instead of using a duration-normalized likelihood or the likelihood ratio, the plain sum of phoneme log-likelihoods $\mathscr{W}_1$ had a higher discriminative ability. Normalization of this feature is possible by dividing with the number of phonemes n in each word

$$(\text{WLh1})\ \mathscr{W}_1 = \sum_{i=1}^{n} L(\vec{X}_i), \quad (\text{WLh2})\ \mathscr{W}_2 = \frac{1}{n}\mathscr{W}_1 \tag{13}$$

The sentence duration score $\mathscr{D}$ is a good word level feature without modifications:

$$(\text{DurS1})\ \mathscr{W}_3 = \sum_{i=1}^{n} S_{\text{dur}}^{(\text{ph})}(T_i * R^{(\text{ph})}|p_i) \tag{14}$$

The following normalizations of $\mathscr{W}_3$ were advantageous in some cases:

$$(\text{DurS2})\ \mathscr{W}_4 = \frac{1}{n}\mathscr{W}_3, \quad (\text{DurS3})\ \mathscr{W}_5 = \mathscr{W}_3\mathscr{R} \tag{15}$$

Confidence measures showed to have the highest discrimination ability. The feature $\mathscr{C}_1$ is a high-level confidence measure derived with the phoneme correlation technique from Cox and Dasmahapatra (2002). It is based on the phoneme confusion matrices for correctly pronounced and mispronounced words. The confusion probabilities are calculated at the frame level. As for the calculation of the likelihood ratio in Eq. (8) q_t^* denotes the phoneme label of the speech frame derived from unconstrained phoneme recognition. The label p_t is obtained from the forced-alignment

$$(\text{PhCfRatio})\ \mathscr{C}_1 = \frac{1}{D}\sum_{t=1}^{D} \log\frac{P(q_t^*|p_t,\text{wrong})}{P(q_t^*|p_t,\text{correct})} \tag{16}$$

Another confidence measure is the word posterior probability (WPP) (Wessel et al., 2001). It measures the degree to which a word recognition hypothesis can be trusted. It may be assumed, that the value of the WPP also reflects the pronunciation quality of a word. The word level pronunciation feature $\mathscr{C}_2$ is based on the sentence likelihood. It was calculated via N-best lists in order to be independent from the architecture and implementation of a speech recognizer

$$(\text{WPP})\ \mathscr{C}_2 = \frac{\sum_{\vec{V}} P(\vec{O}|\vec{V}) f(W_j|V_i)}{\sum_{\vec{V}} P(\vec{O}|\vec{V})} \tag{17}$$

The summation is carried out over the word sequences $\vec{V} = (V_1, V_2, \ldots, V_i, \ldots)$ of each hypothesis from the N-best list. The function $f(W_j|V_i)$ returns 1, if the overlapping condition for the reference word W_j and a word V_i in the hypothesis is met. Otherwise its value is 0. The language model probability $P(\vec{V})$ is not employed for the calculation of the WPP, since the feature should only be based on acoustic evidence.

5.2. Classification method

For the discrimination of the two classes of correctly pronounced and mispronounced words the Gaussian classifier is employed. Other methods for classification, decision trees (CART) and Gaussian mixture models

(GMMs) after reduction of the feature space dimension with principal component analysis (PCA) could not outperform the Gaussian classifier.

6. Experimental setup

Fig. 3 depicts the process of pronunciation feature extraction. The Hidden Markov Model toolkit (Young et al., 2002) is employed for speech recognition. The native acoustic model is trained on about 60 h of American English speech from the WSJ Corpus. There is one context-independent monophone model for 44 English phonemes and one integrated silence and short pause model. Pronunciation features are extracted from forced-alignment and automatic speech recognition results. Phoneme and word recognition were carried out unconstrained, i.e. a statistical language model was not employed.

Phoneme duration statistics are estimated from the TIMIT corpus (Garofolo et al., 1993). Since durations for test sentences cannot be obtained manually and automatically computed durations differ from manual annotation, duration statistics are calculated from the forced-alignment. The phoneme confusion matrices for correctly pronounced and mispronounced words are estimated on the training set of the non-native speech data. The phoneme language model (LM) to calculate the probability of recognized phoneme sequences is estimated on canonic transcriptions of the SI, SA and SX sentences of the TIMIT corpus.

The training and test sets are set up to be disjoint w.r.t. the non-native speakers and the human evaluators. There are four sets which fulfill this condition. Initial experiments are carried out with three sets for training and one set for evaluation. Finally, 4-fold cross-validation is conducted.

In case of linear feature combination, features are selected implicitly by their weighting coefficients. For the Gaussian classifier features selection is possible with floating search. The floating search algorithm works as follows:

(i) Start with the empty feature set $\mathbf{S} = \{\}$.
(ii) Add the relatively best feature to $\mathbf{S}$.
(iii) Remove the relatively worst feature from $\mathbf{S}$, if the quality of the new feature set becomes better than the best set $\mathbf{S}'$ with $|\mathbf{S}| - 1 = |\mathbf{S}'|$ features so far.
(iv) Stop, if a predefined number of features is used, else go to (ii).

$|\mathbf{S}|$ denotes the cardinality of set $\mathbf{S}$. Fig. 4 illustrates the feature selection procedure. As optimization criterion the classification gain defined by the pointwise multiplication of the classifier's confusion matrix $\mathbf{F}$ with a gain matrix $\mathbf{M}$ is employed

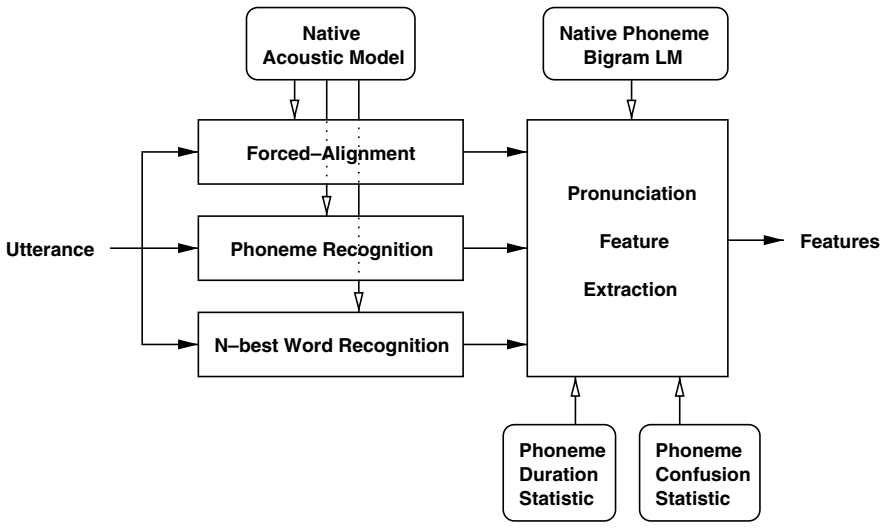

Fig. 3. Experimental setup for pronunciation feature extraction at each level of granularity.

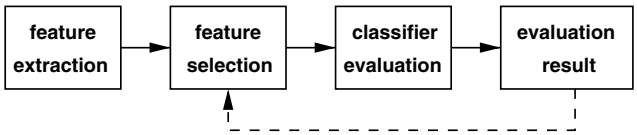

Fig. 4. Feature selection based on floating search.

$$g(\mathbf{M},\mathbf{F}) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k} M_{ij}F_{ij} \tag{18}$$

Here, k is the number of classes, i and j are class indices. The elements F_{ij} of matrix $\mathbf{F}$ denote the probabilities that tokens belonging to class i are classified as class j.

For word classification experiments matrix $\mathbf{M}_1$ is employed. Since classifying a correctly pronounced word as mispronounced might have negative effects on a student than vice versa, the penalty for the former confusion is set higher

$$\mathbf{M}_1 = \begin{bmatrix} +1 & -3 \\ -1 & +1 \end{bmatrix} \tag{19}$$

Matrix $\mathbf{M}_2$ is used for sentence scoring. It is designed so that the more the scoring result deviates from the reference rating, the lower becomes the classification gain

$$\mathbf{M}_2 = \begin{bmatrix} +1 & -1 & -2 & -4 & -8 \\ -1 & +1 & -1 & -2 & -4 \\ -2 & -1 & +1 & -1 & -2 \\ -4 & -2 & -1 & +1 & -1 \\ -8 & -4 & -2 & -1 & +1 \end{bmatrix} \tag{20}$$

The number of samples for each sentence rating as well as the number of correctly pronounced and mispronounced words was highly unbalanced. Resampling was carried out to make the number of samples equal. As many samples as were available for the class with the largest number of samples are selected randomly with replacement for each class.

7. Experimental results

Besides the correlation coefficient (Eq. (1)), the classification gain (Eq. (18)), the class-wise average recognition rate (CL), the average recognition rate tolerating the confusion of neighbored classes (CL-1A), and the total recognition rate (RR) are employed as performance measures. CL is defined as the sum of the diagonal of the classifier's confusion matrix $\mathbf{F}$ divided it by the number of classes. RR measures the percentage of correctly classified items. The quality of pronunciation features is also indicated by their correlation with the human ratings.

7.1. ATR data

7.1.1. Sentence scoring results

The performance for sentence scoring with the Gaussian classifier and linear transformation based on single features is shown in Table 8. The best four single features are the likelihood ratio $\mathscr{K}$ followed by the phoneme accuracy $\mathscr{A}$, the duration score $\mathscr{D}$ and the likelihood score $\mathscr{E}$ normalized by the expected word duration.

By applying the floating search algorithm to one training and test set combination, the feature sets as given in Table 9 ranked by the classification gain are found. The results as given in the tables are obtained using 4-fold cross-validation.

Table 8
Result for sentence scoring based on single pronunciation features (ATR data)

Feature		Gaussian classifier				Linear
ID	Names	Corr.	$g(\mathbf{M}_2, \mathbf{F})$	CL (%)	CL-1A	Corr.
$\mathscr{K}$	(LhRatio)	**0.50**	**−0.35**	**35.5**	**73.0**	**0.51**
$\mathscr{A}$	(PhAcc)	0.44	−0.45	33.9	71.5	0.46
$\mathscr{D}$	(DurScore)	0.42	−0.44	32.8	69.8	0.45
$\mathscr{E}$	(SentLh2)	0.40	−0.47	34.0	70.0	0.44
$\mathscr{L}$	(SentLh1)	0.38	−0.46	31.7	69.5	0.42
$\mathscr{M}$	(PhSeqLh)	0.38	−0.58	32.2	65.1	0.40
$\mathscr{R}$	(MeanPhDur)	0.35	−0.52	30.7	67.1	0.38
$\mathscr{P}$	(PauseDur)	0.31	−0.85	27.1	58.2	0.34

Table 9
Result for sentence scoring with the Gaussian classifier based on multiple pronunciation features (ATR data)

Feature IDs	Feature Names	Corr.	$g(\mathbf{M}_2, \mathbf{F})$	CL (%)	CL-1A
$\mathscr{E}, \mathscr{A}$	(SentLh2,PhAcc)	**0.52**	**−0.26**	36.9	80.0
$\mathscr{E}, \mathscr{A}, \mathscr{R}$	(SentLh2,PhAcc,MeanPhDur)	0.52	−0.27	**38.0**	**80.1**
$\mathscr{E}, \mathscr{A}, \mathscr{R}, \mathscr{M}$	(SentLh2,PhAcc,MeanPhDur,PhSeqLh)	0.52	−0.29	36.6	79.9
$\mathscr{A}, \mathscr{R}, \mathscr{M}, \mathscr{L}, \mathscr{D}$	(PhAcc,MeanPhDur,PhSeqLh,SentLh1,DurScore)	0.52	−0.30	35.2	78.5
$\mathscr{A}, \mathscr{R}, \mathscr{M}, \mathscr{L}, \mathscr{D}, \mathscr{P}$	(PhAcc,MeanPhDur,PhSeqLh,SentLh1,DurScore,PauseDur)	0.51	−0.35	33.8	77.8

The result for linear combination of selected feature sets is given in Table 10. The lower three sets include features which can be calculated given only the forced-alignment. There is a remarkable increase in performance if the features based on the recognition result are also employed (upper three sets).

Considering only the correlation coefficient, the reliability of sentence scoring by linear feature combination is higher than when using the Gaussian classifier. However, the scoring accuracy is much worse, which is obvious when comparing the classification gain. The accuracy is improved by score adjustment with the proposed linear and multiplicative transformation.

Table 11 shows the positive effect of score adjustment. The left matrix is the confusion matrix between reference ratings and rounded scores from the linear classifier. Scoring results '1' or '5' almost never occur. From the right matrix it is clear, that score adjustment contributes remarkably to scoring precision.

Table 10
Result for sentence scoring by linear combination of multiple pronunciation features (ATR data)

Feature IDs	Feature Names	Raw scoring		Score adjustment	
		Corr.	$g(\mathbf{M}_2, \mathbf{F})$	Corr.	$g(\mathbf{M}_2, \mathbf{F})$
$\mathscr{K}, \mathscr{A}, \mathscr{M}, \mathscr{D}$	(LhRatio,PhAcc,PhSeqLh,DurScore)	**0.59**	**−0.66**	**0.59**	**−0.39**
$\mathscr{K}, \mathscr{M}$	(LhRatio,PhSeqLh)	0.56	−0.74	0.58	−0.45
$\mathscr{K}, \mathscr{A}$	(LhRatio,PhAcc)	0.55	−0.70	0.55	−0.44
$\mathscr{D}, \mathscr{E}$	(DurScore,SentLh2)	**0.51**	**−0.73**	**0.52**	**−0.55**
$\mathscr{D}, \mathscr{L}, \mathscr{R}$	(DurScore,SentLh1,MeanPhDur)	0.48	−0.76	0.48	−0.62
$\mathscr{D}, \mathscr{L}$	(DurScore,SentLh1)	0.47	−0.75	0.47	−0.66

Table 11
Confusion matrix obtained after rounding scores from linear classifier with feature combination $\mathscr{K}, \mathscr{A}, \mathscr{M}, \mathscr{D}$ (ATR data)

Reference Rating	Without adjustment					With score adjustment				
	1	2	3	4	5	1	2	3	4	5
1	3	**82**	15	0	0	**52**	35	12	1	0
2	0	**61**	38	1	0	25	**39**	24	10	2
3	0	35	**63**	2	0	11	27	**34**	20	9
4	0	8	**72**	20	0	1	8	21	**31**	**39**
5	0	0	**52**	**48**	0	0	0	4	33	**63**

Table 12 compares the performance of linear feature combination and the Gaussian classifier for native and non-native speech. In case of native speech a reference rating of'1' is assumed for each utterance. The percentage of correctly classified sentences is given by the columns RR. The performance of the Gaussian classifier is superior to linear feature combination. A feature combination with a good scoring accuracy for both native and non-native speakers is (LhRatio) $\mathscr{K}$, (PhAcc) $\mathscr{A}$, (PhSeqLh) $\mathscr{M}$ and (DurScore) $\mathscr{D}$.

7.1.2. Word classification results

Table 13 shows the accuracy for discriminating correctly pronounced words from mispronounced words based on single pronunciation features. The best two features w.r.t. both CL and the classification gain are the phoneme confusion ratio $\mathscr{C}_1$ and the word likelihood $\mathscr{W}_1$.

As for sentence scoring, the floating search algorithm is employed to heuristically find n-best feature sets. The first four combinations in Table 14 are identified when using CL, the last four when using the classification gain as optimization criterion. There was no significant increase in performance if employing five or more features.

Table 12
Result for sentence scoring of non-native and native speech based on multiple pronunciation features (ATR data)

Speech	Feature Names	Native		Non-Native	
Classifier		Linear	Gaussian	Gaussian	
Feature IDs		RR (%)	RR (%)	CL-1A	$g(\mathbf{M}_2,\mathbf{F})$
$\mathscr{K},\mathscr{M}$	(LhRatio,PhSeqLh)	80.1	**90.9**	76.6	−0.30
$\mathscr{K},\mathscr{A},\mathscr{M},\mathscr{D}$	(LhRatio,PhAcc,PhSeqLh,DurScore)	**86.6**	88.5	77.5	−0.30
$\mathscr{D},\mathscr{E}$	(DurScore,SentLh2)	63.7	86.8	75.5	−0.37
$\mathscr{E},\mathscr{A}$	(SentLh2,PhAcc)	84.2	85.0	**80.0**	**−0.26**
$\mathscr{E},\mathscr{A},\mathscr{R},\mathscr{M}$	(SentLh2,PhAcc,MeanPhDur,PhSeqLh)	81.8	83.8	79.9	−0.29

Table 13
Result for word classification with the Gaussian classifier based on single pronunciation features (ATR data)

Feature IDs	Feature Names	CL (%)	$g(\mathbf{M}_2,\mathbf{F})$
$\mathscr{W}_3$	(DurS1)	64.0	**+0.16**
$\mathscr{W}_5$	(DurS3)	61.2	+0.14
$\mathscr{W}_1$	(WLh1)	65.8	+0.07
$\mathscr{C}_1$	(PhCfRatio)	**66.6**	+0.06
$\mathscr{M}$	(PhSeqLh)	64.1	+0.06
$\mathscr{W}_4$	(DurS2)	58.0	+0.02
$\mathscr{C}_2$	(WPP)	66.0	−0.23
$\mathscr{A}$	(PhAcc)	64.7	−0.24
$\mathscr{W}_2$	(WLh2)	54.5	−0.34

Table 14
Result for word classification with the Gaussian classifier based on multiple pronunciation features (ATR data)

Feature IDs	Feature Names	CL (%)	$g(\mathbf{M}_1,\mathbf{F})$
$\mathscr{W}_1,\mathscr{C}_2$	(WLh1,WPP)	70.7	+0.03
$\mathscr{W}_1,\mathscr{C}_2,\mathscr{W}_4$	(WLh1,WPP,DurS2)	71.6	+0.16
$\mathscr{W}_1,\mathscr{C}_2,\mathscr{W}_4,\mathscr{C}_1$	(WLh1,WPP,DurS2,PhCfRatio)	**72.2**	**+0.18**
$\mathscr{W}_1,\mathscr{C}_2,\mathscr{W}_4,\mathscr{C}_1,\mathscr{W}_2$	(WLh1,WPP,DurS2,PhCfRatio,WLh2)	72.1	+0.18
$\mathscr{W}_3,\mathscr{C}_1$	(DurS1,PhCfRatio)	68.3	+0.19
$\mathscr{C}_1,\mathscr{W}_1,\mathscr{W}_5$	(PhCfRatio,WLh1,DurS3)	69.0	+0.23
$\mathscr{C}_1,\mathscr{W}_1,\mathscr{W}_5,\mathscr{A}$	(PhCfRatio,WLh1,DurS3,PhAcc)	**70.4**	**+0.24**
$\mathscr{C}_1,\mathscr{W}_1,\mathscr{W}_4,\mathscr{C}_2,\mathscr{W}_5$	(PhCfRatio,WLh1,DurS2,WPP,DurS3)	69.3	+0.23

In Table 15 the performance with four different feature combinations is compared for native and non-native speech. More than 90% of the words uttered by natives are classified as correctly pronounced. A feature combination with a good scoring accuracy for both native and non-native utterances is (PhCfRatio) $\mathscr{C}_1$, (WLh1) $\mathscr{W}_1$, (DurS3) $\mathscr{W}_5$ and (PhAcc) $\mathscr{A}$.

The reliability of word level mispronunciation detection can be assessed when comparing the confusion matrix of human evaluators with the classifier's confusion matrix. To obtain the former, the majority voting of all but one evaluator was taken as reference and tested against the decision of the remaining evaluator. The average performance of four reference and test combinations is shown in Table 16. There is a disagreement about 8% of the correct words and 42% of the wrong words. From the left table it is clear, that the detection of mispronounced works equally well by automatic classification. However, at the same time the classification error for correctly pronounced words is about 10% higher than for the human evaluators.

The classification accuracy for correct words can be increased at the cost of a decrease in accuracy for mispronounced words. Fig. 5 shows the recall of class "correct" versus recall of class "wrong". The performance of the human evaluators is also indicated. From the graph it is clear, that more than 40% of the mispronounced words are detected while the misclassification error for correctly pronounced words is only about 10%.

Table 15
Comparison of word classification accuracy for native and non-native speech (ATR data)

Feature		Native	Non-native	
IDs	Names	RR	CL (%)	$g(\mathbf{M}_1, \mathbf{F})$
$\mathscr{W}_3$	(DurS1)	**96.2**	64.0	+0.16
$\mathscr{W}_3, \mathscr{C}_1$	(DurS1,PhCfRatio)	93.9	68.3	+0.19
$\mathscr{C}_1, \mathscr{W}_1, \mathscr{W}_5, \mathscr{A}$	(PhCfRatio,WLh1,DurS3,PhAcc)	92.9	**70.4**	**+0.24**
$\mathscr{C}_1, \mathscr{W}_1, \mathscr{W}_5$	(PhCfRatio,WLh1,DurS3)	90.3	69.0	+0.23

Table 16
Comparison of the confusion matrices of the human evaluators and the Gaussian classifier based on the feature set $\{\mathscr{C}_1, \mathscr{W}_1, \mathscr{W}_5, \mathscr{A}\}$ (ATR data)

Machine	Correct	Wrong	Humans	Correct	Wrong
Correct	82.9	17.1	Correct	91.9	8.1
Wrong	42.1	57.9	Wrong	42.2	57.8

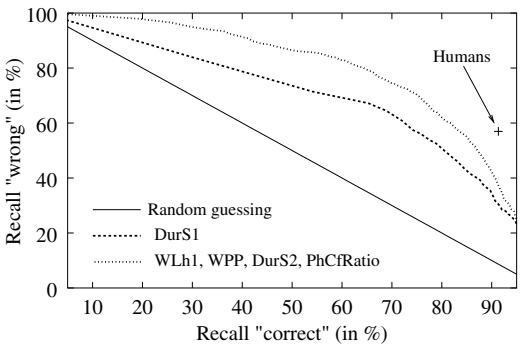

Fig. 5. Recall for both classes of automatic method and performance of human evaluators (ATR data).

7.2. PF-STAR data

7.2.1. Sentence scoring

In order to investigate, whether the pronunciation scoring system works also for speech with different characteristics than its models are trained on, it is evaluated for the PF-STAR non-native database. The acoustic model of the speech recognizer is the same as for experiments with ATR data, and all models of the system are estimated on ATR non-native data.

Table 17 shows the result for sentence scoring. The best performance is achieved with the likelihood ratio score $\mathscr{K}$. There is no improvement, when employing two or more features. In comparison to ATR data, the scoring accuracy is lower w.r.t. all performance measures. Nevertheless, scoring worked reliably enough for at least 68.3% of the utterances, for which at most a confusion with a neighbored rating class occurred. Furthermore, it has to be taken into account, that the PF-STAR sentence ratings are only from one human evaluator.

7.2.2. Word classification

The word scoring performance is only evaluated for feature sets with good results for ATR data. The performance for the best combinations of one to four features are given in Table 19. A class-wise average recognition rate (CL) of 67.4% is achieved with the features (WLh1) $\mathscr{W}_1$ and (WPP) $\mathscr{C}_2$. Table 18 shows the confusion matrices of the human evaluators and the classifier based on this feature combination. Fig. 6 shows the ROC curve of the classifier using this feature combination. The performance of the human evaluators is also indicated.

Although the difference in performance between the automatic method and the humans is larger than for ATR data, results for both sentence scoring and word classification are promising. The feature set shows a high degree of portability despite the fact that training (ATR) and test (PF-STAR) data were collected independently and have different characteristics: children vs. adults, single-accented vs. multi-accented and different text material.

Table 17
Result for sentence scoring with the Gaussian classifier based on single features (PF-STAR data)

Feature IDs	Feature Names	$g(\mathbf{M}_2, \mathbf{F})$	CL (%)	CL-1A
$\mathscr{K}$	(LhRatio)	−0.97	31.8	68.3
$\mathscr{D}$	(DurScore)	−1.46	30.1	61.2

Table 18
Comparison of the confusion matrices of the human evaluators, and the Gaussian classifier based on the feature set $\{\mathscr{W}_1, \mathscr{C}_2\}$ (PF-STAR data)

Machine	Correct	Wrong	Humans	Correct	Wrong
Correct	71.4	28.6	Correct	97.9	2.1
Wrong	37.3	62.7	Wrong	55.6	44.4

Table 19
Result for automatic word classification based on multiple pronunciation features (PF-STAR data)

Feature IDs	Feature Names	CL (%)	$g(\mathbf{M}_1, \mathbf{F})$
$\mathscr{W}_5$	(DurS3)	59.7	+0.07
$\mathscr{W}_1, \mathscr{C}_2$	(WLh1,WPP)	67.4	+0.06
$\mathscr{C}_1, \mathscr{W}_1, \mathscr{W}_5$	(PhCfRatio,WLh1,DurS3)	64.9	+0.11
$\mathscr{W}_1, \mathscr{C}_2, \mathscr{W}_4, \mathscr{C}_1$	(WLh1,WPP,DurS3,PhCfRatio)	66.0	+0.05

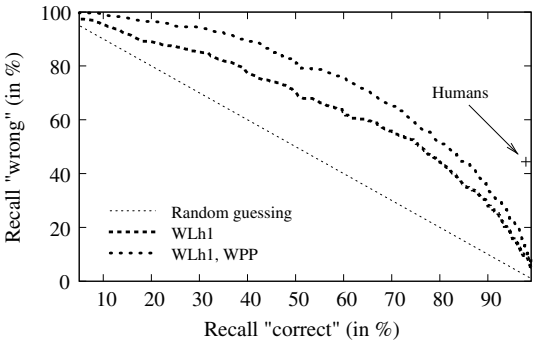

Fig. 6. Recall of both classes with the automatic method and performance of human evaluators (PF-STAR data).

8. Automatic error diagnosis

The purpose of detecting mispronounced tokens in sentences is to identify words which require pronunciation training. If the user desires more detailed feedback, it is possible to obtain information about mispronounced phonemes from a statistic of mispronounced words as depicted in Fig. 7.

In the following, a model to calculate phoneme mispronunciation probabilities given a list of correctly pronounced and mispronounced words is proposed. The approach is data-driven and allows the localization as well as determination of the kind of pronunciation error without knowing the non-native's first language. The model is defined for the phoneme level because speech recognition systems usually work at the phoneme level and to weigh intelligibility higher than strict phonetically correct pronunciation. As notation for English phonemes the SAMPA alphabet (Wells, 1997) is employed.

Given knowledge about the non-native's first language, it is possible to draw further conclusions by comparing the phone sets specified by the International Phonetic Association (IPA, 1999) for the non-native's first language and the target language. For example, if the correct speech sound is not part of the non-native's first language, it is very likely that the learner will have difficulty in producing the sound. If the corresponding sound is in the common phone inventory, however, the mispronunciation is very likely to be due to the learner's unfamiliarity with a word's admissible phoneme sequence.

8.1. Word mispronunciation model

Let $Q = (q_1, \ldots, q_M)$ be the phoneme set of the target language and $P_{\text{mis}}(\vec{w})$ the probability of the event that word $\vec{w}$ consisting of the phoneme sequence $p_1, p_2, \ldots, p_N$ is mispronounced. The mispronunciation probability (MP) of the ith phoneme in a word is denoted by p_i^m, the probability of correct pronunciation by $p_i^c = 1 - p_i^m$. The probability q_j^c can be interpreted as the degree of pronunciation quality of phoneme q_j.

Three possible models for the relationship between the word $P_{\text{mis}}(\vec{w})$ and phoneme MPs p_i^m are given in Table 20. Model (1) assumes, that the probability $P_{\text{mis}}(\vec{w})$ is related to the arithmetic and model (2) to the geometric mean of the phoneme MPs p_i^m. Most elaborate is the Markov chain model (3) given in Fig. 8. It assumes that a word is to be highlighted whenever one or more phonemes are considered mispronounced.

If the phoneme MPs q_j^m were given, the MP of an arbitrary word could be calculated easily. However, the phoneme MP are usually unknown. In the following, a method to estimate the phoneme MPs from word MPs

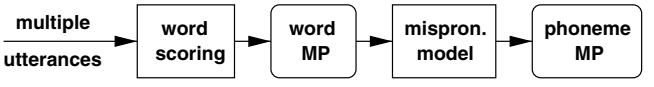

Fig. 7. After scoring a larger number of words, mispronunciation probabilities (MP) of phonemes can be derived from a statistic of mispronounced words.

Table 20
Models for the relationship between the mispronunciation probability (MP) of a word and the MPs of phonemes

Model	Relationship	$P_{\mathrm{mis}}(\vec{w})$
(1)	Geometric mean	$[\prod_{i=1}^{N} p_i^m]^{\frac{1}{N}}$
(2)	Arithmetic mean	$\frac{1}{N}\sum_{i=1}^{N} p_i^m$
(3)	Markov chain	$1.0 - \prod_{i=1}^{N} p_i^c$

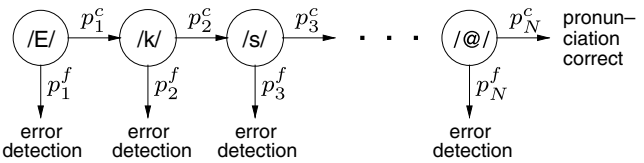

Fig. 8. Mispronunciation and detection process based on a Markov chain. A word is considered mispronounced if one or more of its phonemes are mispronounced and a mispronunciation is always detected by the system.

is described. Given the relationship between phoneme and word MPs as defined by each model in Table 20, it is possible to set up a system of linear equations.

For example, in case of model (2), the equation for the English word "extra" with the phoneme sequence /E k s t r @/ would be

$$P_{\mathrm{mis}}(\mathrm{extra}) = \frac{1}{6}[P_{\mathrm{mis}}(/\mathrm{E}/) + P_{\mathrm{mis}}(/\mathrm{k}/) + \cdots + P_{\mathrm{mis}}(/@/)]$$

or in general

$$P_{\mathrm{mis}}(\vec{w}) = \frac{1}{N}\sum_{j=1}^{M} q_j^m * n_j \tag{21}$$

where n_j is the number of occurrences of phoneme q_j in word $\vec{w}$. The MP of a single word $\vec{w}$ can be obtained by counting its mispronounced vs. all of its occurrences either in the human reference or the automatic word classification result

$$P_{\mathrm{mis}}(\vec{w}) = \frac{\#\text{ times word } \vec{w} \text{ is mispronounced}}{\#\text{ occurrences of word } \vec{w}} \tag{22}$$

If $\vec{x}$ is the vector of phoneme MPs q_j^m for the M target language phonemes, $\vec{b}$ the vector of word MPs $P_{\mathrm{mis}}(\vec{w})$ and the row elements of matrix $\mathbf{A}$ the relative frequencies n_j/N for all words, it is possible to set up a system of the form $\mathbf{A}\vec{x} = \vec{b}$. It only remains to solve the system for $\vec{x}$.

For model (1) it is also possible to obtain a system of linear equations by taking the logarithm on both sides of the equation

$$P_{\mathrm{mis}}(\vec{w}) = \left[\prod_{i=1}^{N} p_i^m\right]^{\frac{1}{N}}$$

i.e. the phoneme MPs will be estimated in log-domain.

Finally, the same approach can also be applied to model (3). It is straightforward to transform the equation

$$P^{(\mathrm{mis})}(\vec{w}) = 1.0 - p_1^c, \ldots, p_N^c = p_1^f + p_1^c p_2^f + \cdots + p_1^c, \ldots, p_{N-1}^c p_N^f$$

defining the relationship between phone and word MPs into

$$\log[1.0 - P^{(\mathrm{mis})}(\vec{w})] = \sum_{i=1}^{N} \log p_i^c = \sum_{j=1}^{M} n_j \log q_j^c$$

which is linear in $\log q_j^c$. Consequently, the rows of matrix $\mathbf{A}$ will be the absolute phoneme frequencies of each word, and the elements of vector $\vec{x}$ the probabilities q_i^c in log-domain.

Usually there are more constraints, i.e. there are more words than phonemes. Consequently, the system will most often be overdetermined and it is unlikely that an error-free solution of the proposed systems exists. However, on the other hand, with too less constraints a unique solution would not exist either. A solution with minimum mean square error between reference and reconstructed word MP can be obtained by solving the system $\mathbf{A}^{\mathrm{T}}\mathbf{A}\vec{x} = \mathbf{A}^{\mathrm{T}}\vec{b}$.

The proposed models are compared by the correlation between the reference and reconstructed word MPs. The reference word MP is obtained from the human annotations. The reconstruction of word MPs from the estimated phoneme MPs is straightforward using the relationships from Table 20.

The evaluation result is shown in Table 21. The correlation is highest for model (3) with a large distance to models (1) and (2). Model (3) can be considered as quite reliable, since the degree of correlation is comparable to the inter-rater correlation. The prediction of phoneme mispronunciation probabilities worked best for Japanese and was most difficult for German.

8.2. Phoneme mispronunciation statistics

8.2.1. ATR data

Table 22 shows the phonemes with the highest mispronunciation probability (MP) for each non-native speaker group when using mispronunciation model (3). A reasonable result are the high MPs for the 'th' sounds /T/, /D/ and the r-colored vowel /3‘/, since the phones [θ], [ð] and [ɝ] are not part of any of the non-native speakers' first language (cf. Table 23). There is also a high MP for /S/ and /@/. For some non-native speaker groups (Chinese, Japanese) this is due to the fact that the phones [ʃ] and [ə] are not part of the non-native's first language. However, mispronunciations may also arise from a speaker's insufficient

Table 21
Comparison of word mispronunciation models by correlation between initially given and reconstructed word MP (ATR data)

Correlation	German	French	Indonesian	Chinese	Japanese
(1) Geometric mean	0.34	0.36	0.40	0.45	0.37
(2) Arithmetic mean	0.41	0.46	0.45	0.49	0.49
(3) Markov chain	**0.55**	**0.64**	**0.67**	**0.65**	**0.69**

Table 22
Ranklist of mispronounced or misread phonemes for each non-native speaker group (ATR data)

German		French		Indonesian		Chinese		Japanese	
T [θ]	0.28	3‘ [ɝ]	0.48	S [ʃ]	0.57	T [θ]	0.51	T [θ]	0.47
3‘ [ɝ]	0.26	T [θ]	0.40	3‘ [ɝ]	0.47	S [ʃ]	0.49	3‘ [ɝ]	0.46
j [j]	0.25	@ [ə]	0.30	T [θ]	0.38	D [ð]	0.34	S [ʃ]	0.35
S [ʃ]	0.25	D [ð]	0.28	tS [ʧ]	0.30	3‘ [ɝ]	0.34	@ [ə]	0.29
aU[aʊ]	0.21	aU[aʊ]	0.25	@ [ə]	0.30	aU[aʊ]	0.30	r [ɹ]	0.27
g [g]	0.17	V [ʌ]	0.24	j [j]	0.27	r [ɹ]	0.29	l [l]	0.27

Table 23
Consonants and vowels of American English which are missing in each of the non-native's first language

Group	Missing consonants	Missing vowels
German	[θ][ð][ɹ][w]	[ɚ][ɝ][æ][ʌ][ɑ]
French	[θ][ð][ɹ][ŋ][h][ʤ]	[ɚ][ɝ][ɪ][æ][ʌ][ɑ][ʊ]
Indonesian	[θ][ð][ɹ][ʒ]	[ɚ][ɝ][ɪ][æ][ʌ][ɛ][ʊ][ɔ]
Chinese	[θ][ð][ɹ][ʒ][ʃ][v][z][b][d][g][ʤ]	[ɚ][ɝ][æ][ʌ][ɑ][e]
Japanese	[θ][ð][ɹ][ʒ][ʃ][v][f][l]	[ɚ][ɝ][ɪ][æ][ʌ][ɑ][ɛ] [ʊ][ɔ][ə]

knowledge of a word's correct pronunciation, e.g. for 'extra', 'box' and 'bugle' (cf. Table 24). A further interesting result is the high MP of phoneme /r/ for Chinese, presumably realized as [l], and of phonemes /r/ and /l/ for Japanese, since there is only a phoneme realized as [ɾ] in Japanese.

The statistic of Table 22 was estimated using the human annotations. However, no human reference will be available in case of an automatic system. Consequently, it has to be investigated whether a useful mispronunciation statistic can be derived from the automatic word scoring result. Table 25 shows the phonemes with the highest MP when using the word MPs either calculated from the human reference or from the automatic word scoring result. The feature set (PhCfRatio) $\mathscr{C}_1$, (WLh1) $\mathscr{W}_1$, (DurS2) $\mathscr{W}_4$, (WPP) $\mathscr{C}_2$, (DurS3) $\mathscr{W}_5$ was employed for automatic word classification. For German, French, Chinese and Japanese, the statistics have three out of five and for Indonesian they have all phonemes in common. The automatic detection of pronunciation errors which are common among all speaker groups regarding the 'th' sounds [θ][ð] and the r-colored vowels [ɚ][ɝ] appears to be most reliable (cf. Table 24).

Further insight into the nature of mispronunciation errors can be gained from a phoneme confusion statistic. Such a statistic can be obtained by aligning each sentence's canonical phoneme transcription with the automatic transcription from phoneme recognition constrained with a bigram phoneme LM. It is straightforward to calculate a phoneme confusion probability matrix from the phoneme level alignment. Table 26 shows the phonemes ranked by their confusion probability for each first language group. It is obvious, that speech sounds which are not part of the non-native's first language (•) have most often a high confusion probability.

Table 24
Words with a high mispronunciation probability calculated from the human reference (ATR data)

Word	Pronunciation	All	German	French	Indonesian	Chinese	Japanese
Extra	[ɛkstɹə]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Exposure	[ɪkspoʃɚ]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Exam	[ɪgzæm]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Box	[bɑks]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Mirage	[məɹɑʃ]	**0.92**	0.71	**1.00**	**0.94**	**0.94**	**0.92**
Centrifuge	[sɛntɹɪfjʊdʒ]	**0.85**	**0.93**	**0.86**	**0.94**	**0.89**	0.72
Bugle	[bjʊgəl]	**0.85**	0.64	**1.00**	**1.00**	**0.89**	0.72
Frantically	[fræntɪkli]	**0.84**	0.79	**0.81**	**0.88**	0.67	**0.96**
Purchase	[pɝtʃəs]	0.76	0.64	**0.94**	**0.81**	0.67	0.76
Rare	[ɹɛɹ]	0.75	0.36	0.69	0.75	**0.89**	**0.88**
Contagious	[kəntedʒəs]	0.74	0.57	0.69	**0.81**	**0.83**	0.72
Formula	[fɔɹm j ələ]	0.73	0.79	**0.88**	**0.81**	0.67	0.56
Ambulance	[æmbjələns]	0.73	0.64	**0.81**	0.75	0.78	0.72
Development	[dɪvɛləpmənt]	0.70	0.36	**0.88**	**0.94**	0.67	0.64
Pizzerias	[pitsɚiəz]	0.69	0.36	0.56	**0.88**	0.78	0.76
Guard	[gɑɹd]	0.69	0.43	0.75	0.75	**0.83**	0.68
Colored	[kʌlɚd]	0.69	0.50	**0.81**	0.56	0.67	**0.84**
Chablis	[ʃəbli]	0.69	0.36	0.44	**0.88**	**0.83**	**0.80**
Thursdays	[θɝzdez]	0.68	0.50	0.69	0.75	0.67	0.76
Mergers	[mɝdʒɚz]	0.67	0.29	0.69	**0.81**	0.72	0.72

Table 25
Frequently misread or mispronounced phonemes either derived using the human word level annotations or the automatic word scoring result (ATR data)

Group		Human reference		Automatic word scoring
German	[θ] [ɝ] [j] [ʃ] [aU]	/T/,/3‘/,/j/,/S/,/aU/	[ɝ] [θ] [ɚ] [aU] [ə]	/3‘/,/T/,/@‘/,/aU/,/@/
French	[ɝ] [θ] [ə] [ð] [aʊ]	/3‘/,/T/,/@/,/D/,/aU/	[ɝ] [θ] [ə] [ɚ] [ʊ]	/3‘/,/T/,/@/,/@‘/,/U/
Indonesian	[ʃ] [ɝ] [θ] [ʧ] [ə]	/S/,/3‘/,/T/,/tS/,/@/	[ɝ] [θ] [ʃ] [ə] [ʧ]	/3‘/,/T/,/S/,/@/,/tS/
Chinese	[θ] [ʃ] [ð] [ɝ] [aʊ] [ɹ]	/T/,/S/,/D/,/3‘/,/aU/,/r/	[θ] [ɝ] [l] [j] [ʃ]	/T/,/3‘/,/l/,/j/,/S/
Japanese	[θ] [ɝ] [ʃ] [ɚ] [ɹ] [l]	/T/,/3‘/,/S/,/@‘/,/r/,/l/	[ɝ] [θ] [ə] [ɚ] [aʊ]	/3‘/,/T/,/@/,/@‘/,/aU/

Moreover, Table 27 shows a selection of phone substitutions derived from this phoneme confusion matrix for the phoneme mispronunciation statistic from Table 25. The most common mistake for at least four non-native speaker groups are the confusions [θ][s], [ʃ][s], and the confusion of the r-colored vowel [ɚ] with [ə][ɛ]. The confusion of [aʊ] with [ɔ][o] is present for three speaker groups. In case of Japanese natives, the confusion of [ɪ][l] and the confusion of [ə] with [o][i] is a reasonable result, since there are only five vowels realized as [a], [ɯ], [e], [i] and [o] in Japanese.

8.2.2. PF-STAR data

Finally, it is investigated, what kind of pronunciation mistakes are most common for German children reading English sentences from their textbook (PF-STAR data). Estimation of phone mispronunciation probabilities is carried out for the human reference as well as the scoring result. All phonemes except diphthongs

Table 26
Phoneme confusion probability ranking obtained from forced-alignment and phoneme recognition for each non-native accent (ATR data)

Rk	German	French	Indonesian	Chinese	Japanese
1	•/@ˈ/[ɚ]	•/3ˈ/[ɝ]	•/3ˈ/[ɝ]	•/T/[θ]	•/3ˈ/[ɝ]
2	•/3ˈ/[ɝ]	•/@ˈ/[ɚ]	•/T/[θ]	•/A/[ɑ]	•/@ˈ/[ɚ]
3	•/A/[ɑ]	•/A/[ɑ]	/tS/[ʧ]	•/@ˈ/[ɚ]	•/A/[ɑ]
4	/U/[ʊ]	/aU/[aʊ]	•/U/[ʊ]	/U/[ʊ]	/aU/[aʊ]
5	/aU/[aʊ]	•/I/[ɪ]	/v/[v]	•/3ˈ/[ɝ]	•/U/[ʊ]
6	/I/[ɪ]	•/T/[θ]	/A/[ɑ]	/I/[ɪ]	•/I/[ɪ]
7	/d/[d]	•/U/[ʊ]	/dZ/[dʒ]	/l/[l]	•/T/[θ]
8	•/T/[θ]	•/D/[ð]	/z/[z]	/aU/[aʊ]	•/v/[v]
9	/z/[z]	/l/[l]	•/@ˈ/[ɚ]	•/v/[v]	/dZ/[dʒ]
10	/v/[v]	/v/[v]	/aU/[aʊ]	•/D/[ð]	•/l/[l]
11	•/D/[ð]	•/dZ/[dʒ]	•/D/[ð]	•/z/[z]	•/D/[ð]
12	•/{/[æ]	•/V/[ʌ]	•/I/[ɪ]	•/d/[d]	•/@/[ə]
13	/dZ/[dʒ]	/E/[ɛ]	/S/[ʃ]	•/@/[ə]	•/E/[ɛ]
14	/l/[l]	/O/[ɔ]	•/V/[ʌ]	•/V/[ʌ]	•/r/[ɹ]
15	/O/[ɔ]	/z/[z]	/d/[d]	•/S/[ʃ]	/z/[z]
16	/tS/[ʧ]	/@/[ə]	/@/[ə]	•/g/[g]	•/{/[æ]
17	/g/[g]	/d/[d]	•/{/[æ]	•/dZ/[dʒ]	•/S/[ʃ]
18	•/V/[ʌ]	•/r/[ɹ]	•/r/[ɹ]	•/{/[æ]	/d/[d]
19	/@/[ə]	•/{/[æ]	/g/[g]	•/r/[ɹ]	/tS/[ʧ]
20	/E/[ɛ]	•/tS/[ʧ]	•/O/[ɔ]	/N/[ŋ]	/g/[g]
21	•/r/[ɹ]	/S/[ʃ]	•/E/[ɛ]	/tS/[ʧ]	/O/[ɔ]
22	/t/[t]	/g/[g]	/l/[l]	/O/[ɔ]	•/V/[ʌ]
23	/N/[ŋ]	/j/[j]	/t/[t]	/I/[ɪ]	/w/[w]
24	/j/[j]	/o/[o]	/p/[p]	/j/[j]	/j/[j]
25	/I/[ɪ]	/N/[ŋ]	/e/[e]	/e/[e]	•/I/[ɪ]
26	/S/[ʃ]	/I/[ɪ]	/w/[w]	/E/[ɛ]	•/f/[f]
27	/u/[u]	/e/[e]	/j/[j]	/t/[t]	/t/[t]
28	/i/[i]	/i/[i]	/o/[o]	/w/[w]	/N/[ŋ]
29	•/w/[w]	/t/[t]	/I/[ɪ]	•/b/[b]	/o/[o]
30	/aI/[aɪ]	/u/[u]	/N/[ŋ]	/i/[i]	/b/[b]
31	/e/[e]	/m/[m]	/i/[i]	/u/[u]	/m/[m]
32	/h/[h]	/b/[b]	/u/[u]	/o/[o]	/n/[n]
33	/m/[m]	/p/[p]	/aI/[aɪ]	/aI/[aɪ]	/p/[p]
34	/b/[b]	/w/[w]	/b/[b]	/n/[n]	/aI/[aɪ]
35	/f/[f]	/aI/[aɪ]	/m/[m]	/h/[h]	/e/[e]
36	/p/[p]	/f/[f]	/f/[f]	/m/[m]	/h/[h]
37	/o/[o]	•/h/[h]	/h/[h]	/p/[p]	/u/[u]
38	/n/[n]	/n/[n]	/k/[k]	/s/[s]	/s/[s]
39	/s/[s]	/s/[s]	/s/[s]	/f/[f]	/i/[i]
40	/k/[k]	/k/[k]	/n/[n]	/k/[k]	/k/[k]
41	/OI/[ɔɪ]	/Z/[ʒ]	/OI/[ɔɪ]	/OI/[ɔɪ]	•/Z/[ʒ]
42	/Z/[ʒ]	/OI/[ɔɪ]	•/Z/[ʒ]	•/Z/[ʒ]	/OI/[ɔɪ]

Table 27
Phone substitutions derived from forced-alignment and phoneme recognition (ATR data)

German	French	Indonesian	Chinese	Japanese
[θ] → [s]	[ɝ] → [ɛ]	[ʃ] → [s]	[θ] → [s][z]	[θ] → [s][t]
[ɝ] → [o]	[θ] → [s][t]	[ɝ] → [ɛ]	[ʃ] → [s]	[ɝ] → [ɔ]
[j] → [i][ɪ]	[ə] → [ɪ][o]	[θ] → [t][s]	[ð] → [d]	[ʃ] → [s]
[ʃ] → [s]	[ð] → [d]	[ʧ] → [t]	[ɝ] → [o][ʊ]	[ɚ] → [ɛ]
[aʊ] → [o][ɔ]	[aʊ] → [o][ɔ]	[ə] → [o]	[aʊ] → [ɔ][o]	[ɹ][l] → [l][ɹ]
[ɚ] → [ə]	[ɚ] → [ɛ][ə]	[ɚ] → [ɛ][ə]	[r] → [l]	[ə] → [o][i]

Table 28
Frequently misread or mispronounced phonemes derived from the human annotations and the automatic word scoring result (PF-STAR data)

Human reference		Automatic word scoring	
0.18	/v/[v]	0.25	/dZ/[dʒ] → /S/[ʃ], /tS/[ʧ]
0.16	/dZ/[dʒ]	0.25	/T/[θ] → /D/[ð], /s/[s]
0.15	/w/[w]	0.21	/S/[ʃ] → /Z/[ʒ]
0.12	/D/[ð]	0.19	/Z/[ʒ] → /S/[ʃ]
0.11	/T/[θ]	0.18	/3/[ɜ] → /V/[ʌ]
0.11	/@/[ə]	0.15	/g/[g] → /k/[k]

are considered. Phoneme confusion pairs for phonemes with a high MP are extracted from the phoneme confusion matrix. The result is given in Table 28.

Among the top mispronunciation candidates derived from the word level human annotations are the phonemes /w/, /D/ and /T/. This is a reasonable finding, since the corresponding speech sounds [w][ɝ][θ] are not part of the German phone inventory. The automatic statistic has two phonemes in common with the reference statistic among the top five candidates, /dZ/ and /T/. Further phonemes with a high MP are /S/ and /Z/.

These examples show that based on the automatically derived phoneme mispronunciation and confusion statistics, the automatic system would in principle not only be able to provide feedback on mispronounced phonemes over time, but could also assemble automatically special training sessions for phonemes the foreign language student often mispronounces.

9. Conclusion

This paper proposes a method for pronunciation scoring, which is independent from the student's first language and can in principle be applied to other target languages. Besides investigating features and methods for scoring words and sentences, an approach to automatic diagnosis of phoneme mispronunciations based on word scoring results is presented.

In case of the multi-accent, adult speaker, non-native English ATR database, a likelihood ratio score, the phoneme recognition accuracy, the phoneme sequence probability and a duration score were the most successful feature combination for scoring sentences. The word posterior probability and phoneme confusion probability ratio of correctly pronounced and mispronounced words are identified as new word level features. These two confidence measures together with word likelihood and phoneme recognition accuracy were the best feature combination.

Scoring the pronunciation quality of sentences was as reliable as the human evaluation. Although scores obtained by linear feature combination had a higher correlation with the human reference, a more accurate result was obtained with a single-density Gaussian classifier. Furthermore, it is shown that the scoring accuracy of the linear classifier can be improved when applying a linear and polynomial transformation.

The likelihood ratio and the duration features had the highest portability when scoring non-native speech of German children in the PF-STAR database given a system trained on ATR data only. The same is true for

the word level. The word likelihood and word posterior probability were most portable. Possible reasons for the discrepancy are the different data characteristics, adult vs. children, single vs. multi-accented and different text material.

Promising results for detecting mispronounced words in non-native speech are achieved for both databases: a class-wise average recognition rate of 72% for the ATR and 67% for the PF-STAR non-native speech data. It was also shown, that about 90% of the words uttered by natives are classified as correctly pronounced even if classifier parameters are trained on non-native data only. Perfect detection of mispronounced words (and phonemes) remains difficult, however, since there is even disagreement about mispronounced items among human evaluators.

In order to gain more insights about the nature of pronunciation mistakes, a data-driven approach to automatically derive a statistic of mispronounced phonemes from mispronounced words is presented. The relationship between phoneme and word mispronunciation probabilities was investigated by three models. The model best fitting the data was a Markov chain model, which assumes that a word is mispronounced whenever one or more of its phonemes is mispronounced. Reasonable results for five non-native accent groups and both databases are obtained without using any prior knowledge on the first language of the non-native speakers. The automatically derived information on mispronounced words and phonemes can serve as valuable feedback to the learner or enable the system to assemble special training sessions.

Acknowledgement

This research was supported in part by the National Institute of Information and Communications Technology (NICT), Japan.

References

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russel, M., Steidl, S., Wong, M., 2005. The PF-STAR children's speech corpus. In: Proceedings of the European Conference on Speech, Communication and Technology (Eurospeech), pp. 2761–2764.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M., 1990. Automatic evaluation and training in english pronunciation. In: Proceedings of the ICSLP, pp. 1185–1188.

Cox, S., Dasmahapatra, S., 2002. High-level approaches to confidence estimation in speech recognition. IEEE Transactions on Speech and Audio Processing 10 (7), 460–471.

Cucchiarini, C., Strik, H., Binnenpoorte, D., Boves, L., 2000. Pronunciation evaluation in read an spontaneous speech: a comparison between human ratings and automatic scores. In: Proceedings of the New Sounds.

Franco, H., Neumeyer, L., Digalakis, V., Ronen, O., 2000. Combination of machine scores for automatic grading of pronunciation quality. Speech Communication 30, 121–130.

Garofolo, J., et al., 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Technical Report, Philadelphia.

Gruhn, R., Cincarek, T., Nakamura, S., 2004. A multi-accent non-native English database. In: Proceedings of the Acoustical Society of Japan, pp. 163–164.

Herron, D., Menzel, W., Atwell, E., Bisiani, R., Daneluzzi, F., Morton, R., Schmidt, J., 1999. Automatic localization and diagnosis of pronunciation errors for second-language learners of English. In: Proceedings of the European Conference on Speech, Communication and Technology (Eurospeech), pp. 855–858.

IPA, 1999. Handbook of the International Phonetic Association. Cambridge University Press.

Ito, A., Lim, Y.-L., Suzuki, M., Makino, S., 2005. Pronunciation error detection method based on error rule clustering using a decision tree. In: Proceedings of the European Conference on Speech, Communication and Technology (Eurospeech), pp. 173–176.

Minematsu, N., 2004. Yet another acoustic representation of speech sounds. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 1669–1672.

Moustroufas, N., Digalakis, V., 2007. Automatic pronunciation evaluation of foreign speakers using unknown text. Computer, Speech and Language 21 (1), 219–230.

Neri, A., Cucchiarini, C., Strik, H., 2002a. Feedback in computer assisted pronunciation training: technology push or demand pull? In: Proceedings of the International Conference on Spoken Language Processing, pp. 1209–1212.

Neri, A., Cucchiarini, C., Strik, H., Boves, L., 2002b. The pedagogy-technology interface in computer assisted pronunciation training. Computer Assisted Language Learning 15, 441–447.

Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M., 2000. Automatic scoring of pronunciation quality. Speech Communication 30, 83–93.

Park, J.G., Rhee, S.-C., 2004. Development of the knowledge-based spoken english evaluation system and it's application. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP).

Teixeira, C., Franco, H., Shriberg, E., Precoda, K., Sönmez, K., 2000. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In: Proceedings of the International Conference on Spoken Language Processing, vol. 3, pp. 187–190.

Tsubota, Y., Kawahara, T., Dantsuji, M., 2002. CALL system for Japanese students of English using pronunciation error prediction and formant structure estimation. In: Proceedings of the Conference on Integration of Speech Technology into Learning (INSTIL).

Wells, J., 1997. SAMPA: A Computer Readable Phonetic Alphabet. <http://www.phon.ucl.ac.uk/home/sampa/home.htm>.

Wessel, F., Schlüter, R., Macherey, K., Ney, H., 2001. Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing 9 (3), 288–298.

Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication 30, 95–108.

Young, S., Everermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. HTK Speech Recognition Toolkit. <http://htk.eng.cam.ac.uk/>.