Pronunciation Feature Extraction

Christian Hacker[1,⋆], Tobias Cincarek[2,⋆⋆], Rainer Gruhn[2,⋆⋆⋆], Stefan Steidl[1], Elmar Nöth[1], and Heinrich Niemann[1]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Martensstraße 3, D-91058 Erlangen, Germany
[2] ATR Spoken Language Translation Res. Labs., Kyoto, Japan
hacker@informatik.uni-erlangen.de

Abstract. Automatic pronunciation scoring makes novel applications for computer assisted language learning possible. In this paper we concentrate on the feature extraction. A relatively large feature vector with 28 sentence- and 33 word-level features has been designed. On the word-level correctly and mispronounced words are classified, on the sentence-level utterances are rated with 5 discrete marks. The features are evaluated on two databases with non-native adults' and children's speech, respectively. Up to 72 % class-wise-averaged recognition rate is achieved for 2 classes; the result of the 5-class problem can be interpreted as 80 % recognition rate.

1 Introduction

Pronunciation scoring is the automatic assessment of the pronunciation quality of phonemes, words, utterances, or larger units especially for non-native speakers. A possible application are systems for computer assisted pronunciation training (CAPT) to support the student of a foreign language to acquire correct pronunciation. In this paper a set of 28 sentence-level features is proposed which encodes a high amount of information that is important to grade the pronunciation of a sentence. A similar set of 33 features has been developed to reject mispronounced words. Some simple features that highly correlate with human marking are e.g. the word or phone recognition rate obtained by an automatic speech recognition system.

As reference two databases are applied and compared: The ATR/SLT NON-NATIVE-database, recorded at the Spoken Language Translation Research Laboratories (SLT) of ATR [5], contains speech of non-native adults from different countries reading English phrases. Secondly, the PF-STAR NON-NATIVE-database is applied. In the European project PF-STAR (http://pfstar.itc.it/)

⋆ 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.

⋆⋆ Now with Graduate School of Information Science, Nara Institute of Science and Technology, Japan; a part of this work was funded by the National Institute of Information and Communication Technology (NICT), Japan.

⋆⋆⋆ Now with Temic SDS, Ulm, Germany.

W. Kropatsch, R. Sablatnig, and A. Hanbury (Eds.): DAGM 2005, LNCS 3663, pp. 141–148, 2005.

native and non-native children's speech has been recorded from English, Italian, Swedish, and German partners. In this paper the data of German children reading English texts is investigated. For both databases human ratings are available for word- and sentence-level. Our feature set for automatic grading of the non-natives' English has been developed for the ATR/SLT NON-NATIVE data set. Although the age of the speakers, the recording conditions, the speakers' English proficiency, and the instruction of the labelers are different for both databases good classification results are achieved for the PF-STAR NON-NATIVE data, too.

2 Related Work

Neumeyer et al. [7] automatically score non-natives on the sentence- and speaker-level. The inter-rater open-correlation is 0.78 (sentence) and 0.87 (speaker). Correlations with different machine scores (likelihood, posterior scores, accuracy, duration and syllabic timing) are calculated. Different combination techniques for sentence based marks are investigated in [4]: with neural networks a correlation of 0.64 is achieved. Different aspects of human rating and different machine scores are compared in [2]. For phone-level scoring the liklihood based *Goodness of Pronunciation* measure is analyzed in [9,10]. In [6] a novel phonological representation of speech is used to grade the pronunciation.

3 Corpora

The pronunciation features will be evaluated on two different databases:

The part of the ATR/SLT NON-NATIVE-database [5] used in the following consists of 6.4 hours of speech: the 96 non-native speakers (81m, 15f, age 21 – 52) were reading 48 phonetically rich sentences from the TIMIT SX set with a vocabulary of 395 words. The first language of most speakers is Japanese, Chinese, German, French or Indonesian. Each speaker read each sentence usually only once. However, he was asked to repeat the recording of a sentence, when he completely misread or forgot to utter a word or made too long pauses between words. Further, a repetition was possible, if the speaker was not satisfied with the recorded utterance. 15 English teachers (native speakers) evaluated the data: Each utterance has been marked by 3 – 4 teachers, each teacher marked 24 speakers. They assigned a sentence-level rating from 1 (best) to 5 (worst) in terms of pronunciation and fluency and marked any mispronounced words.

The PF-STAR NON-NATIVE-database contains 3.4 hours of speech from 57 German children (26m, 31f, age 10 – 15) reading English texts, recorded by the University of Erlangen. Most children had been learning English for half a year only. They were reading known texts from their text book and some phrases and single words, which also have been recorded by our partners in the Pf-Star project. The recordings contain reading errors, repetitions of words, word fragments and nonverbals. The total size of the vocabulary is 940 words. A German student of English (graduate level) marked mispronounced words and rated the data on sentence-level. Further markings of mispronounced words by 12 teachers will be available soon.

To train speech recognizers, that are needed for the pronunciation feature extraction, further databases are applied. Read speech from the Wall Street Journal (Wsj)[1]-corpus is used to train an adults' recognizer for the Atr/Slt data; the Pf-Star Native-database contains read speech from British English children [3] and is used to train the recognizer for the Pf-Star Non-Native children. Some phone statistics are estimated from the Timit[2]-database.

4 Input Data for Pronunciation Feature Extraction

In our scenario the candidate who is practicing English is reading known texts. Classification of the candidate's pronunciation quality is performed in feature space. Our feature extraction requires several outputs of a speech recognizer and some statistics, which are explained in the following.

Word and phone recognition: The HTK toolkit is used for the estimation of monophone models and for the decoding. 39 features are extracted every 10 ms: 12 cepstral coefficients and the normalized log-energy with first and second derivatives. Cepstral mean subtraction is applied. The number of codebook mixtures was increased successively during training until 16 mixtures were reached. 44 3-state phoneme HMMs and silence models were retrained for four iterations after each mixture increment. The acoustic models for the non-native adults are built with native English data from the Wsj-corpus. The recognizer is evaluated on the Hub2 evaluation test set from Wsj. With a bigram language model (LM) 80.8 % word accuracy (WA) are achieved. The children's speech recognizer is trained with the Pf-Star Native data. With a bigram LM 40.8 % WA are achieved on the native testing data-set.
Native phoneme language model: To compute prior probabilities of phone sequences obtained by unconstrained phoneme recognition, a bigram phoneme LM will be employed. The LM is estimated from the Timit-corpus.
Native phoneme duration statistic: In order to calculate the expected duration of words and phones or to estimate posterior probabilities of an observed length of time, the distribution of phoneme durations has to be modeled. They are estimated on the Timit-database after forced-alignment.
Phone confusion matrices: A Phone confusion occurs, if the reference phone and the recognized phone differ. Phone confusion matrices are estimated separately for both, the correctly pronounced and the mispronounced words. These matrices contain the probabilities $P(q|p)$, that phoneme p is recognized as q. The confusion matrices are estimated on the Atr/Slt Non-Native-corpus. The reference sequence is obtained by forced-alignment and the recognized sequence is obtained with the phone-recognizer trained on Wsj.

5 Pronunciation Features

Next, a set of 28 sentence based pronunciation features is described, that is an extension of the features in [7]. After this 33 word-level features are proposed.

[1] http://www.ldc.upenn.edu/, catalog number LDC93S6
[2] http://www.ldc.upenn.edu/, catalog number LDC93S1

Table 1. *Correlation between each of the 28 sentence-features and the human rating (*ATR/SLT *data).*

R1	*R2*	*R3*	*R4*	*R5*	*P1*	*P2*	*D1*	*D2*	*DS1*	*DS2*	*L1*	*L2*	*L3*
−.34	−.37	+.37	+.39	−.32	+.33	+.32	+.30	+.28	−.45	−.46	−.24	−.34	−.28
L4	*L5*	*L6*	*L7*	*L8*	*L9*	*LR1*	*LR2*	*LR3*	*A1*	*A2*	*PS1*	*PS2*	*PS3*
−.41	−.42	−.37	−.35	−.41	−.43	−.48	−.50	−.52	−.45	−.38	−.22	−.28	−.40

5.1 Sentence-Level Features

First, some notations, that will be referred to in the following: Let t_i be the duration of phone number i in the utterance (pauses are not counted), and T_s the duration of the sentence. We introduce $T = \sum_{i=1}^{n} t_i \leq T_s$ as the sum of phone-durations per sentence without pauses. Assume further m to be the number of words and n the number of phones per sentence. Then the rate-of-speech is defined as

$$R^{(phon)} = n/T_s \quad \text{or} \quad R^{(word)} = m/T_s \tag{1}$$

To evaluate the features proposed in the following the correlation between automatic scores and human rating is analyzed for the ATR/SLT NON-NATIVE-database. The reference is the mean of the marks of the different human raters. An overview of features and correlation-values can be found in Tab. 1. Particularly since for the PF-STAR data only one rater is available, correlation coefficients would be clearly lower. Eight feature categories are built from the set of 33 sentence-level pronunciation features. Within such a feature-set elements differ mainly in the way of normalization.

Rate-Of-Speech *(R)*: This category comprises 5 components: $R^{(word)}$ and $R^{(phon)}$ referred to as *R1* and *R2*, both reciprocals (*R3*, *R4*) and the phonation time ratio T/T_s. If we compare the features with the human annotation, absolute correlations between 0.32 and 0.39 are obtained. The best feature is *R4*.
Pauses *(P)*: The total duration of between-word pauses (*P1*) is correlated with sentence-level ratings by 0.33. Normalization of the pause duration by the number of pauses did not lead to an increase of correlation. The number of between word pauses longer than 0.2 sec. (*P2*) correlates with 0.32.
DurationLUT *(D)*: Elements of this category are computed from the duration statistics (look-up-table, LUT) introduced in Sect. 4. For all phonemes the expected duration d_i from the LUT is used to compute the deviation $|t_i - d_i|$. *D1* is the mean duration deviation, *D2* the scatter. The correlation with the human annotation is 0.30 and 0.28. In other feature groups d_i is used for normalization.
DurationScore*(DS)*: Phoneme duration statistics have been estimated on native data (Sect. 4). To calculate *DS*-features for non-natives, we first normalize the observed phoneme duration (obtained by forced alignment) with the rate-of-speech; we achieve $\bar{t}_i$. Using natives' statistics we now calculate the probability $\log P(\bar{t}_i|p, \boldsymbol{x})$ given the phone p in the reference and the acoustic observation $\boldsymbol{x}$. Summing up these probabilities of an utterance *DS1* is achieved. After normalization with n (*DS2*) the correlation with the reference rating is -0.46.

Likelihood *(L)*: This category contains 9 features based on log-likelihood scores $L(\boldsymbol{x}) = \log P(\boldsymbol{x}|\lambda_q)$ of the acoustic observation $\boldsymbol{x}$ given the HMM λ of the phone q the decoder has elected. The sentence likelihood *L1* can be approximated by the sum of all phoneme log-likelihoods if independence of the phones is assumed. By normalizing with n or m we obtain *L2* and *L3*. The global and local sentence likelihood as introduced in Neumeyer et al. [7] is additionally normalized by $R^{(phon)}$:

$$L4 = \frac{1}{R^{(phon)}} \frac{\sum_{i=1}^{n} L(\boldsymbol{x})}{T} \qquad L5 = \frac{1}{nR^{(phon)}} \sum_{i=1}^{n} \frac{L(\boldsymbol{x})}{t_i} \tag{2}$$

L6 and *L7* are based on word-likelihoods. First, we normalize and then we average per sentence. By further replacing the observed phone duration t_i with d_i (from the duration statistic LUT) we get *L8* and *L9* from *L5* and *L6*. Best correlation with the human reference is achieved with *L9* (-0.43) and *L5* (-0.42).
LikeliRatio *(LR)* comprises features that compare the likelihoods received from the forced alignment and the phone recognizer; in log-space for each frame both values are subtracted and summed up over the entire utterance. For *LR1* we normalize with n, for *LR2* with $T \cdot R^{(phon)}$ and for *LR3* with $\sum_{i=1}^{n} d_i R^{(phon)}$. Correlation with human annotations is around 0.5.
Accuracy *(A)*: Human ratings and the phoneme or word accuracy (*A1* and *A2*) correlate with -0.45 and -0.38. Since a sentence contains only few words, the phone recognition rate can be calculated more robustly.
PhoneSeq *(PS)*: With a phoneme bigram LM estimated on native-data (Sect. 4), the a priori probability $\log P(\boldsymbol{q}|LM)$ of the observed phone sequence $\boldsymbol{q}$ can be computed (*PS1*). After normalization with n or the rate-of-speech *PS2* and *PS3* are obtained. The latter correlates -0.40 with human marks.

5.2 Word-Level Features

On word-level 33 features, partly similar to the sentence-features, are extracted from the data. *Pauses* and *LikeliRatio* are not considered. Here, *Rate-of-Speech*-features are based on the number of phonemes per word duration. The category *DurationLUT* contains amongst others the expected word duration, which is the sum of expected phone durations from the native duration statistic (Sect. 4). As for the *DurationScore*, the phone duration probabilities are now summed up for each word. The *Likelihood* group comprises features with similar normalizations as discussed above. Additionally minimum, maximum and scatter of frame-based log-likelihood values are taken into account. *Accuracy* only contains the phone accuracy. Given a phoneme bigram LM, the probability of the phone sequence corresponding to the current word is calculated in *PhoneSeq*. Additionally we compute the following features:

PhoneConfusion*(PC).* Instead of *LikeliRatio*-features *PC*-features are calculated on the word-level. Both groups compare forced alignment and phone-recognition. Phone-confusion occurs, if the reference phone p and the recognized phone q differ. From the two precalculated confusion matrices (Sect. 4), we get the probabilities $P(q|p)$ given either the class *wrongly pronounced* or the class

Table 2. *Word-/sentence-level classification with 1D-features. CL in %*

	word (2 classes)		sentence (3 cl.)		sentence (5 cl.)	
	PF-STAR	ATR/SLT	PF-STAR	ATR/SLT	PF-STAR	ATR/SLT
Rate-Of-Speech	59.5	65.9	41.2	54.5	28.6	32.2
Pauses	–	–	41.4	48.0	**29.1**	30.7
DurationLUT	58.3	64.5	40.0	54.7	23.7	35.0
DurationScore	62.1	**67.3**	**44.9**	**55.9**	28.5	**37.6**
Likelihood	62.6	**67.0**	41.1	**56.7**	26.4	**35.6**
LikeliRatio	–	–	**44.0**	**61.8**	27.0	**41.9**
PhoneConfusion	**65.5**	65.6	–	–	–	–
Accuracy	**64.6**	61.5	**47.8**	52.0	**29.1**	34.9
PhoneSeq	59.4	65.0	42.4	52.8	28.6	34.9
Confidence	61.5	**67.2**	–	–	–	–
Context	53.1	51.8	–	–	–	–

correctly pronounced. For each frame, the ratio of both probabilities is computed; the mean *(PC1)*, maximum *(PC2)*, minimum *(PC3)*, scatter *(PC4)*, and median *(PC5)* are used as features.

Confidence*(CF).* We measure with 3 *CF*-features the probability of words in the reference sequence, given a non-native's utterance. The assumption is: the better the pronunciation of a particular word, the higher is its posterior probability. The calculation of the word posteriors is based on n-best lists.

Context*(C)*-features are obtained by comparing word and sentence based likelihood scores or by calculating the fluctuation of either the local rate-of-speech or the local duration ratio between expected and observed word duration (7 features). Let $R_j^{(local)}$ be the number of phones per word duration of the j-th word, then the fluctuation *C2* is

$$C2 = \frac{2R_j^{(local)}}{R_{j-1}^{(local)} + R_{j+1}^{(local)}} \tag{3}$$

6 Results

To evaluate the pronunciation features we applied the leave-one-speaker-out cross-validation approach. We use the LDA-classifier and, additionally, for the experiments in the last paragraph the Gaussian classifier. For all experiments the class-wise averaged recognition rate (CL)[3] and in some cases additionally the overall recognition rate (RR) is given. Tab. 2 shows classification results for the individual feature components. For each feature category, the optimal result is shown as well for the word-level (2 classes: correctly pronounced / mispronounced) as for the sentence-level. On the sentence-level classes of neighboring marks overlap clearly, thus the classification results are rather low. For the 3-class

[3] Average of recalls (not weighted by prior probabilities). For unbalanced data robust recognition is required for both, classes with many and classes with few elements.

Table 3. *Results for 2 databases and 2 classification levels*

	word (CL in %) 2 classes	sentence (RR in %) 5 classes ±1 tolerance
PF-STAR cross-vali, all features	69.1	72.9
Training: ATR/SLT, test: PF-STAR	67.7	61.9
ATR/SLT cross-vali, feature selection	72.2	79.9

task we map mark 2 → 1 and 4 → 3. The reference rating is for the ATR/SLT NON-NATIVE data as follows: a word is considered to be mispronounced, if it is marked by at least 2 raters. On sentence-level, the discrete mean of the different teachers' marks is calculated. The PF-STAR data is on both levels marked by one rater, only. Consequently labels are less robust and the recognition results lower. Further reasons for the lower recognition rate on PF-STAR is, that the recognition of children's speech seems to be more difficult [8], that the utterances contain reading errors and word fragments, and that the overlap between training and test is smaller (the phone-confusion matrices are estimated from the ATR/SLT training-set; the ATR/SLT-speakers read TIMIT sentences as used for the phoneme LM and duration statistics). Best features on word-level are *Accuracy* and *PhoneConfusion* for the children's data and *DurationScore*, *Likelihood*, and *Confidence* for the adults. On sentence-level we obtained good results for PF-STAR with *Accuracy*, *DurationScore*, and *LikeliRatio*, for ATR/SLT in particular with *LikeliRatio*.

On the PF-STAR-corpus we investigate whether recognition rates increase if the entire feature-set is employed. Again we use the LDA-classifier. On word-level with 33 features 69.1 % CL (72.0 % RR, Tab. 3) are achieved. Features are highly correlated, nevertheless we gain 3.6 % points in comparison to the best single feature. On sentence-level best results are achieved after reduction of the 28 features to 14 principal-components, since otherwise not enough training data would be available. For the 3-class task CL is 50.2 % (52.6 % RR), for the 5-class task 33.4 % (28.7 % RR). If we allow confusion of neighboring marks, e.g. classifying mark 2 as 1, the recognition rate can be interpreted as 72.9 % RR (Tab. 3). Fortunately, the pronunciation features are transferable between different corpora: we train classifiers with the ATR/SLT-corpus and test them with PF-STAR children. On word-level 67.7 % CL are achieved; tolerating 1 mark deviation the sentence-level result can be interpreted as 61.9 % RR.

Further investigation were conducted with the Gaussian classifier and the floating search feature-selection algorithm using the ATR/SLT-corpus [1]. One optimal combination with five word-level features comprises features from the categories *Context*(2), *Confidence*, *Likelihood*, and *DurationScore*: 72.2 % CL are achieved (Tab. 3). On sentence-level 40.1 % CL for five classes is derived from *Accuracy*, *Likelihood* and *LikeliRatio*; with the best single feature (*LikeliRatio*) 36.7 % are achieved using Gauss and 41.9 % using LDA (cf. Tab. 2). If we allow the confusion of neighboring marks, the recognition rate can be interpreted as 79.9 % RR (Tab. 3). Further, assuming natives to have perfect pronunciation, they are recognized with 90.2 % RR using the 5-class recognizer.

7 Conclusion

In this paper two non-native speech databases with children's (PF-STAR) and adults' speech (ATR/SLT) are described. Raters marked correct and mispronounced words and graded the sentences with marks 1 – 5. For the adult's data ratings from 3 – 4 native teachers are available, for the children's data only one rating of a student of English. We described a set of 28 sentence-based pronunciation features and 35 word-level features. Best correlation with human ratings is obtained with the *LikeliRatio*-features, which compare the log-likelihood of forced-alignment and recognized phone-sequence. For classification experiments we employ leave-one-speaker-out cross-validation approach. With single features we get recognition rates up to 67 % (2 classes, word-level), 62 % (3 classes, sentence-level) and 42 % (5 classes, sentence-level). Due to the less precise rating, higher variability of children's speech, and the fact that the children's corpus contains reading errors and word fragments, worse results are achieved for the PF-STAR data. By combining features the recognition could be increased. With feature selection a combination of features could be found, that includes separately not well performing features like *Context*-features, that seem to contain additional information. Further could be shown that the features are transferable: After training with ATR/SLT data, we evaluated with PF-STAR data and obtained acceptable results. If we evaluate natives, in deed 90 % are recognized as very good speakers. For future work we expect further improvement from the combination of both classification levels.

References

1. T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura. Pronunciation Scoring and Extraction of Mispronounced Words for Non-Native Speech. In *Proc. Acoustical Society of Japan*, pages 141–142, 2004.
2. 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 Communication*, 30:109–119, 2000.
3. S.M. D'Arcy, L.P. Wong, and M.J. Russell. Recognition of Read and Spontaneous Children's Speech Using two New Corpora. In *Proc. ICSLP*, Korea, 2004.
4. H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communication*, 30:121–130, 2000.
5. R. Gruhn, T. Cincarek, and S. Nakamura. A Multi-Accent Non-Native English Database. In *Proc. of the Acoustical Society of Japan*, 2004.
6. N. Minematsu. Pronunciation Assessment Based upon Phonological Distortions Observed in Language Learners' Utterances. In *Proc. ICSLP*, Korea, 2004.
7. L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. Automatic Scoring of Pronunciation Quality. *Speech Communication*, 30:83–93, 2000.
8. G. Stemmer, C. Hacker, S. Steidl, and E. Nöth. Acoustic Normalization of Children's Speech. In *Proc. Eurospeech*, pages 1313–1316, Geneva, Switzerland, 2003.
9. S.M. Witt and S.J. Young. Language Learning Based on Non-Native Speech Recognition. In *Proc. Eurospeech*, pages 633 – 636, Rhodes, Greece, 1997.
10. S.M. Witt and S.J. Young. Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communication*, 30:95–108, 2000.