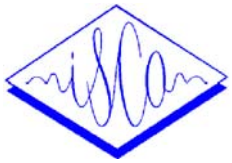

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AUTOMATIC DETECTION OF PHONE-LEVEL MISPRONUNCIATION FOR LANGUAGE LEARNING

Horacio Franco, Leonardo Neumeyer, María Ramos, and Harry Bratt

SRI International
Speech Technology and Research Laboratory
Menlo Park, CA, 94025, USA

ABSTRACT

We are interested in automatically detecting specific phone segments that have been mispronounced by a nonnative student of a foreign language. The phone-level information allows a language instruction system to provide the student with feedback about specific pronunciation mistakes. Two approaches were evaluated; in the first approach, log-posterior probability-based scores [1] are computed for each phone segment. These probabilities are based on acoustic models of native speech. The second approach uses a phonetically labeled nonnative speech database to train two different acoustic models for each phone: one model is trained with the acceptable, or correct native-like pronunciations, while the other model is trained with the incorrect, strongly nonnative pronunciations. For each phone segment, a log-likelihood ratio score is computed using the incorrect and correct pronunciation models. Either type of score is compared with a phone dependent threshold to detect a mispronunciation. Performance of both approaches was evaluated in a phonetically transcribed database of 130,000 phones uttered in continuous speech sentences by 206 nonnative speakers.

1. INTRODUCTION

Computer-based language instruction systems potentially can offer some advantages over traditional methods, especially in areas such as pronunciation training, which often require full attention of the teacher to a single student. If the computer could provide the type of feedback that a pronunciation teacher provides, it would be a much cheaper alternative, accessible at any time and at any place, and certainly tireless.

Recent advances in research on automatic pronunciation scoring [1],[2] allow us to obtain pronunciation quality ratings for sentences or groups of sentences, with arbitrary text, with grading consistency similar to that of an expert teacher. While pronunciation scoring is essential in systems designed for automatic evaluation, a score or number represents only part of the desired feedback for language instruction. In the classroom, a human teacher can point to specific problems in producing the new sounds, and can give specific directions to lessen the most salient pronunciation problems. Our current efforts focus on the introduction of detailed feedback on specific pronunciation problems to help correct or improve pronunciation.

Native and nonnative pronunciations differ in many dimensions. For example, at the phone-segment level, there are differences in phonetic features, which lie on a continuum of possible values between L1 and L2 [3]. There are also prosodic elements, such as stress, duration, timing, pauses, and intonation, which are crucial to native-like pronunciation [4], although in this work we are focusing on segmental pronunciation aspects.

To provide useful feedback at the phone-segment level we need to reliably detect whether a phone is native-like or nonnative, and, ideally, to evaluate how close it is to the native phone production along its different phonetic features. Recently, the use of posterior scores was extended to evaluate the pronunciation quality of specific phone segments [5] as well as to detect phone mispronunciations [6],[7]. An alternative approach [8] used hidden Markov models (HMMs) with two alternative pronunciations per phone—one trained on native speech, the other on strongly nonnative speech. Mispronunciations were detected from the phone backtrace when the nonnative phone alternative was chosen.

The recent availability of a large nonnative speech database [9] with detailed phone-level transcriptions allowed us to accurately extend and evaluate our phone mispronunciation detection strategies. We investigated two different methods for the detection of phone-level mispronunciations (rather than scoring the pronunciation quality of phone segments). In the first method, posterior scores [1] are computed for each phone segment. These probabilities are based on acoustic models of native speech. The second method uses the phonetically transcribed nonnative speech database to train two different acoustic models for each phone: one model is trained with the acceptable, or correct native-like pronunciations, while the other model is trained with the incorrect, strongly nonnative pronunciations. For each phone segment, a log-likelihood ratio score is computed using the incorrect and correct pronunciation models. Either type of score is compared with a phone dependent threshold to detect a mispronunciation. Both methods were evaluated over a phonetically transcribed database of 130,000 phones uttered by 206 nonnative speakers of Latin American Spanish.

2. DATABASE DESCRIPTION

The collection of phone-level pronunciation data is one of the most challenging tasks necessary to build and evaluate a system that can give detailed feedback on specific phone-level pronunciation problems. For this study we had collected a Latin-American Spanish speech database [9] that included recordings from native and nonnative speakers. A group of expert phoneticians transcribed part of the nonnative speech data to be used for development of the mispronunciation detection algorithms. This effort involved obtaining detailed phone-level information for approximately 130,000 phones uttered in continuous speech sentences. These sentences were produced by 206 nonnative speakers whose native language was American English. Their levels of proficiency were varied, and an attempt was made to balance the number of speakers by level

of proficiency as well as by gender. For this study, the detailed phone-level transcriptions were collapsed into two categories: native-like and nonnative pronunciations. In this way we conveyed the judgment of the nativeness of each phone occurrence.

Four native Spanish-speaking phoneticians provided the detailed phonetic transcriptions for 2550 sentences totalling 130,000 phones that were randomly divided among transcribers. An additional 160 sentences, the common pool, were transcribed by all four phoneticians to assess the consistency with which the human could achieve this task. In [9] it was concluded that not all the phone classes could be transcribed consistently. The most reliable to transcribe were the approximants /β/, /δ/, and /γ/; surprisingly, some of the phones which were expected to be good predictors of nonnativeness, such as voiceless stops, most vowels, and /l/ and /r/, did not have good consistency across all the transcribers.

3. DETECTION OF MISPRONUNCIATION

The mispronunciation detection is planned to be integrated into our existing pronunciation scoring paradigm [2],[1], which uses an HMM speech recognizer to generate phonetic segmentations of the student's speech. The recognizer models can be trained using a database of native speakers or a combination of native and nonnative speakers. From the alignments, different pronunciation scores can be generated for each phonetic segment, using different types of models. The scores of the different phonetic segments are combined and calibrated to obtain the closest match to the judgment of an expert human rater. We typically assume that the interaction between the student and the computer has been designed to be error-free. In this case the phonetic segmentation can be obtained by computing a forced alignment using the known prompted text and the pronunciation dictionary.

The goal of the phone-level mispronunciation detection is to add a judgment of nativeness for each phonetic segment defined in the forced alignment.

3.1 Definition of the Mispronunciation Labels

To evaluate, as well as to train, the models used in the mispronunciation detection algorithms, we need to define for each phone segment whether or not it was pronounced in a native-like manner. To this end, we define the canonical transcription as the sequence of "expected" phones; this phone string is obtained from the recognizer forced alignment.

To assess what was actually uttered, we associated the canonical transcription with the transcriber's phone string by applying a dynamic programing (DP) alignment of the two strings. The distance between phone labels was based on the actual acoustic distance between phone classes. This type of distance allowed us to disambiguate phone insertions and deletions in the mapping of the strings. Then, each phone in the canonical transcription was assigned a label "correct" or "mispronounced", depending on whether or not the transcriptions from the phoneticians were the same as the canonical transcription. Phone segments labeled as "correct" correspond to a native-like phone. Phone segments labeled as "mispronounced" may correspond to a nonnative version of the same phone, a different phone, or a deletion of the phone. By our definition, phone insertions induce a "mispronounced" label for the canonical phone to which they are mapped.

Informal analysis has shown that the recognizer forced alignments are robust to the variability of the nonnative pronunciations. On the other hand, phone insertions and deletions may induce some minor alignment errors. This problem could be alleviated by adding more alternative pronunciations to the pronunciation dictionary in order to model the most common deletions and insertions.

3.2 Human Detection of Mispronunciations

To assess an overall measure of consistency across transcribers, we aligned the transcription from each phonetician with the canonical transcription as described above. From these alignments we derived the sequence of "correct" and "mispronounced" labels for the phones of each sentence. We then compared the sequence of labels between every pair of raters by counting the percent of the time that they disagree. The average across all the pairs of raters is an estimate of the mean disagreement between human raters. The resulting value of 19.8% can be considered as a lower bound to the average detection error that an automatic detection system may achieve.

3.3 Mispronunciation Detection Methods

Two approaches were evaluated and compared. Both assume that a phonetic segmentation of the utterance has been obtained in a first step by using the Viterbi algorithm and the known transcription.

In the first approach, previously developed log-posterior probability-based scores [1] are computed for each phone segment with canonic label q_i. For each frame the posterior probability $P(q_i|y_t)$ of the phone q_i given the observation vector y_t is

$$P(q_i|y_t) = \frac{p(y_t|q_i)P(q_i)}{\sum_{j=1}^{M} p(y_t|q_j)P(q_j)} \quad (1)$$

The sum over j runs over a set of context-independent models for all phone classes. $P(q_i)$ is the prior probability of the phone class q_i. The posterior score $\rho(q_i)$ for the phone segment is defined as

$$\rho(q_i) = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log P(q_i|y_t) \quad (2)$$

where d is the frame duration of the phone and t_0 is the starting frame index of the phone segment. The class conditional phone distributions $p(y_t|q_i)$ used to compute the posterior probabilities are Gaussian mixture models that have been trained with a large database of *native* speech. For a mispronunciation to be detected, the phone posterior score $\rho(q_i)$ must be below a threshold predetermined for each phone class.

The second approach uses the phonetically labeled nonnative database to train two different Gaussian mixture models for each phone class: one model is trained with the "correct", native-like pronunciations of a phone, while the other model is trained with the "mispronounced" or nonnative pronunciations of the same phone. A four-way jackknifing procedure was used to train and evaluate this approach on the same phonetically transcribed nonnative database. There were no common speakers across the evaluation and training sets. The number of

Gaussians per model was proportional to the amount of training data for each model, ranging from 200 to 1.

In the evaluation phase, for each phone segment q_i, a length-normalized log-likelihood ratio $LLR(q_i)$ score is computed for the phone segment by using the "mispronounced" and "correct" pronunciation models λ_M and λ_C, respectively.

$$LLR(q_i) = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} [\log p(y_t|q_i, \lambda_M) - \log p(y_t|q_i, \lambda_C)] \quad (3)$$

The normalization by d allows definition of unique thresholds for the LLR for each phone class, independent of the lengths of the segments. A mispronunciation is detected when the LLR is above a predetermined threshold, specific to each phone.

For both detection approaches, and for each phone class, different types of performance measures were computed for a wide range of thresholds, receiver operating characteristic (ROC) curves were obtained, and optimal thresholds were determined for the points of equal error rate (EER).

4. EXPERIMENTS

The acoustic models used to generate the phonetic alignments and produce the posterior scores were gender independent, Genonic Gaussian mixture models, as introduced in [10]. They were trained using a gender-balanced database of 142 native Latin American Spanish speakers, totaling about 32,000 sentences. Given the alignments, the detection of mispronunciation is reduced to a binary decision problem as the phone class is given by the alignments. Consequently, the mispronunciation detection performance can be studied for each phone class independently. Reasons to evaluate the performance for each phone class are that (1) the distributions of machine scores corresponding to different phone classes have different statistics, so independent thresholds must be used for each phone class, (2) the percent of "correct" and "mispronounced" reference labels is different for each phone class, and (3) the complexity of the acoustic model may be different for each phone class.

The performance of the mispronunciation detection algorithms was evaluated as a function of the threshold, for each phone class. For each threshold we obtained the machine-produced labels "correct" (C) or "mispronounced" (M) for each phone utterance. Then, we compared the machine labels with the labels obtained from the human phoneticians. Two types of performance measures were computed for each threshold: error measures and correlation measures.

The error measures were the total error, which is the percent of cases where the machine label and the human label differ; the probability of false detection, estimated as the percent of cases where a phone utterance is labeled by the machine as incorrect when it was in fact correct; and the probability of missing a target, that is, the probability that the machine labeled a phone utterance as correct when it was in fact incorrect.

To compute the correlation measures we first converted the C and M labels to numeric values 0 and 1, respectively. Then the 0-1 strings from machine and human judgments for each phone were correlated using the standard correlation coefficient, as well as the cross correlation measure (or cosine distance) used in [7].

One important issue we found is that for many phone classes the number of phone utterances that have been labeled "mispronounced" by the phoneticians was much less than the number labeled "correct". The probability of error for those phone classes was then very biased by the priors. This bias, combined with the fact that the distributions of machine scores for "correct" and for "mispronounced" phones had significant overlap for some phone classes, determined that in some cases the probability of error could be relatively low, when we would just be classifying every phone utterance as correct, regardless of its machine score. For that reason the minimum error point was not a good indicator of how well we can actually detect a mispronunciation. In addition, in comparing detection performance across phone classes, the measure should not be affected by the priors of the labels. Consequently, we evaluated the mispronunciation performance by computing the ROC curve, and finding the points of EER, where the probability of false detection is equal to the probability of missing a target. This error measure is independent of the actual priors for the C or M labels, but results in a higher total error rate when the priors are skewed.

To some degree, a similar but complementary effect was observed for the cross correlation measure; that is, for the phone classes with relatively high detection error rate, relatively high cross correlation values could be obtained by just labeling every phone utterance as mispronounced.

In Table 1 we show the EER, the correlation coefficient, and the cross correlation measure for each phone class and for both detection methods. Weighted averages overall all the phones are also shown. The phones whose nativeness or mispronunciation were detected most reliably were the approximants /β/, /ð/, and /γ/, the voiced stops /b/ and /d/, the fricative /x/, and the semivowel /w/. These phone classes have good agreement with those found to be the most consistent across different transcribers [9]. The LLR method performed better than the posterior-based method for almost all phone classes. The lower performance for the nasals /m/ and /ñ/ could be explained because they had very few training examples for the mispronounced phones. The reduction in error rate was not uniform across phone classes. The advantage of the LLR method over the posterior method is more significant if we look only at the phone classes with the lowest detection error. On average, the EER had a relative reduction of 33% for the seven most reliably detected phone classes referred above, when going from posterior-based to LLR-based detection. Acceptable levels of the correlation coefficients were also found for that set of phones.

The overall weighted average of the phone mispronunciation detection EER was 35.5% when log-posterior scores were used while 32.3% EER was obtained when the LLR method was used. If instead of the EER, we obtain the minimum total detection error for each phone class, and compute the average error weighted by the number of examples in each class, the resulting minimum average error is 21.3% for the posterior-based method and 19.4% for the LLR-based method. This minimum average error can be compared with the transcribers' percent of pairwise disagreement reported in section 3.2, as both take into account the actual priors of the evaluation data. The closeness of the human-machine and the human-human average errors suggests that the accuracy of the LLR-based detection method is bounded by the consistency of the human transcriptions.

Phone	Posterior score			LLR score		
	EER	Corr	Cross	EER	Corr	Cross
a	35.0	0.28	0.35	35.6	0.29	0.44
b	29.8	0.41	0.76	15.5	0.72	0.88
β	29.8	0.39	0.80	21.5	0.57	0.87
c	44.8	0.34	0.43	39.6	0.28	0.48
d	34.1	0.35	0.75	20.3	0.61	0.82
δ	26.6	0.48	0.75	19.0	0.63	0.84
e	37.9	0.29	0.40	37.4	0.29	0.50
f	27.9	0.20	0.20	27.9	0.27	0.29
g	41.8	0.18	0.46	28.8	0.40	0.55
γ	29.1	0.42	0.84	20.1	0.59	0.88
i	28.2	0.42	0.52	26.2	0.45	0.57
k	38.3	0.26	0.59	32.4	0.37	0.70
l	29.3	0.38	0.56	28.9	0.43	0.60
m	25.0	0.26	0.27	28.2	0.29	0.31
n	41.8	0.15	0.41	35.4	0.23	0.43
ñ	33.3	0.54	0.60	45.1	0.18	0.41
o	38.2	0.22	0.31	39.4	0.20	0.46
p	41.9	0.17	0.47	32.9	0.35	0.64
r	38.2	0.24	0.55	34.5	0.33	0.68
rr	35.3	0.26	0.76	33.7	0.31	0.88
s	36.7	0.19	0.29	27.8	0.35	0.47
t	37.8	0.25	0.45	31.1	0.41	0.62
u	33.8	0.36	0.47	35.3	0.32	0.47
w	23.5	0.53	0.72	14.9	0.74	0.85
x	16.7	0.61	0.69	15.1	0.75	0.79
y	32.5	0.34	0.41	32.3	0.33	0.48
z	44.6	0.13	0.72	30.7	0.48	0.80
Avg.	35.5	0.27	0.45	32.3	0.34	0.54

Table 1: Equal error rate, human-machine correlation, and cross correlation at the phone level for the two detection methods studied. Weighted averages of the various scoring measures are shown at the bottom.

5. DISCUSSION AND CONCLUSIONS

We studied two mispronunciation detection algorithms. One algorithm is based on posterior probability scores computed using models of the native speech, and the other is based on models trained on actual nonnative speech, including both "correct" and "mispronounced" phone utterances.

An important advantage of the posterior-based method is that the native models can be applied to detect mispronunciation errors for any type of nonnative accent. The LLR-based method needs to be trained with specific examples of the target nonnative user population. Experimental results show that the LLR-based system has better overall performance than the posterior-based method. The improvement is particularly significant for the phone classes with the highest consistency across transcribers.

The results also suggest that the reported performance of the system might have been limited by the accuracy and consistency of the transcriptions. This is suggested by (1) the agreement between the most consistent phone classes for humans and the best recognized phone classes by the machine, (2) the similar level of average error rate between pairs of humans on one hand and between machine and humans on the other hand, and (3) the fact that the level of improvement, when using the LLR method, is more significant on the most consistent phone classes.

Results showed the set of phones where mispronunciation can be detected reliably. They mostly coincide with those phones that the phoneticians were able to transcribe more consistently. The overall error rate of the best system was 19.4%, which was similar to an estimate of pairwise human disagreement on the same task.

6. ACKNOWLEDGMENTS

Special thanks to Mitchel Weintraub and Françoise Beaufays for valuable help with the model training software. We gratefully acknowledge support from the U. S. Government under the Technology Reinvestment Program (TRP). The views expressed in this material do not necessarily reflect those of the Government.

REFERENCES

[1] H. Franco, L. Neumeyer, and Y. Kim (1997), Automatic Pronunciation Scoring for Language Instruction, *Proc. of ICASSP 97*, pp. 1471-1474, Munich.

[2] L. Neumeyer, H. Franco, M. Weintraub, and P. Price (1996), Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech, *Proc. of ICSLP 96*, pp. 1457-1460, Philadelphia, Pennsylvania.

[3] J. Flege (1980), Phonetic Approximation in Second Language Acquisition, *Language Learning*, Vol. 30, No. 1, pp. 117-134.

[4] M. Eskenazi (1996), Detection of Foreign Speakers' Pronunciation Errors for Second Language Training - Preliminary Results, *Proc. of ICSLP 96*, pp. 1465-1468.

[5] Y. Kim, H. Franco, and L. Neumeyer (1997), Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction, *Proc. of EUROSPEECH 97*, pp. 649-652, Rhodes.

[6] S. Witt and S. Young (1997), Language Learning Based on Non-Native Speech Recognition, in *Proc. of EUROSPEECH 97*, pp. 633-636, Rhodes.

[7] S. Witt and S. Young (1998), Performance Measures for Phone-Level Pronunciation Teaching in Call, *Proc. of the Workshop on Speech Technology in Language Learning*, pp. 99-102, Marholmen, Sweden.

[8] O. Ronen, L. Neumeyer, and H. Franco (1997), Automatic Detection of Mispronunciation for Language Instruction, *Proc. of EUROSPEECH 97*, pp. 645-648, Rhodes.

[9] H. Bratt, L. Neumeyer, E. Shriberg, and H. Franco (1998), Collection and Detailed Transcription of a Speech database for Development of Language Learning Technologies, *Proc. of ICSLP 98*, pp. 1539-1542, Sydney, Australia.

[10] V. Digalakis and H. Murveit (1994), GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer, *Proc. of ICASSP 94*, pp. I537-I540.