AUTOMATIC DETECTION OF MISPRONUNCIATION FOR LANGUAGE INSTRUCTION

Orith Ronen, Leonardo Neumeyer, and Horacio Franco

Speech Technology and Research Laboratory
SRI International, Menlo Park, California 94025 USA
http://www.speech.sri.com

ABSTRACT

This work is part of a project aimed at developing a speech recognition system for language instruction that can assess the quality of pronunciation, identify pronunciation problems, and provide the student with accurate feedback about specific mistakes. Previous work was mainly concerned with scoring the quality of pronunciation. In this work we focus on automatic detection of mispronunciation. While scoring quantifies the mispronunciation, detection identifies the occurrence of a specific problem. Detecting pronunciation problems is necessary for providing feedback to the student. We use pronunciation scoring techniques to evaluate the performance of our mispronunciation model.

1. INTRODUCTION

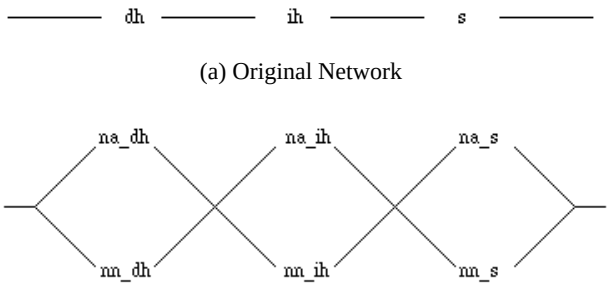
This work presents a method for detecting mispronunciations as a part of a computer-based language instruction system [1, 2]. The goal is to develop a speech recognition system for language instruction, that can assess the quality of pronunciation, identify pronunciation problems and provide the student with accurate feedback about specific mistakes.

The basic pronunciation scoring paradigm previously developed [3, 4, 5] uses hidden Markov models (HMMs) [6] to generate phonetic segmentations of the student's speech. From these segmentations, machine scores are obtained based on HMM log-likelihoods, phone durations, HMM phone posterior probabilities [1, 7], and a combination of these scores [7]. The scores are computed using native acoustic models, and they represent the degree of match between the nonnative speech and the native models. The effectiveness of each machine score is evaluated based on its correlation with human scores on a large database of nonnative speakers. The best result was obtained using average phone segment posterior probability, with a correlation of $r = 0.58$ at the sentence level and $r = 0.88$ at the speaker level. The level of human-machine correlation for the phone posterior probability score was comparable to the duration measure for the case of overall sentence scoring. Moreover, it performed significantly better than both likelihood and duration scores for the case of phone-specific scoring. Using score combination, we improved the sentence-level correlation to $r = 0.62$ [7].

In this work, we investigate techniques for detecting mispronunciations rather than scoring the quality of a given segment. We developed a mispronunciation model allowing us to detect phones with nonnative pronunciation, that is, mispronounced. To do this, we not only model the native but also the nonnative speech data. This approach can be enhanced by incorporating knowledge about the expected set of mispronunciations for a given pair of languages.

Given the subjective nature of the problem, one of our main concerns is to validate the results by using human judgments. Therefore, we collected a large database of human ratings of overall pronunciation quality, as described below, but have only a limited number of human ratings for specific phone segments. To take advantage of the large number of overall pronunciation ratings, we first evaluate the performance of the proposed mispronunciation model by generating pronunciation scores using the model and computing the correlation between them and the human scores. We also compare the performance of the new pronunciation scoring techniques with previous techniques.

The database of pronunciation quality we use in this work consists of speech from 100 natives of Parisian French (native data) and from 100 American students speaking French (nonnative data). A panel of five French teachers rated the overall pronunciation of each nonnative sentence on a scale of 1 to 5 ranging from unintelligible to native quality. These human ratings are the reference for the sentence-level scoring obtained by the machine, and the average human score per speaker is the reference for speaker-level evaluation. For a small subset of the database, the raters also scored the pronunciation quality of specific phone segments. A companion paper [8] presents the evaluation of automatic pronunciation scoring of specific phone segments.



(a) Original Network

(b) Mispronunciation (MP) Network

Figure 1. Two pronunciation networks for the word 'this': (a) the original network with a single pronunciation for each phone; (b) the mispronunciation network where each phone has two alternative pronunciations: native ('na_') and nonnative ('nn_').

2. MODELING MISPRONUNCIATION

In our previous work, our approach was to model speech as a sequence of phone HMMs trained with native speech data only, as shown in Figure 1.a. In this paper we expand the model to be a network with alternate pronunciations. Each phone in the network can optionally be pronounced either as a native or as a nonnative, as shown in Figure 1.b. We will refer to this graph as the mispronunciation (MP) network.

Native phone models are initialized using a subset of the native training speech data. Nonnative phone models are initialized using the subset of the nonnative data that was scored low (in the range of 1 to 2) by the human raters. The training procedure is as follows. We duplicate each entry in the dictionary by assigning a native and a nonnative variant to each word. The pronunciation of native words consists of a linear sequence of native phones. The pronunciation of nonnative words uses the MP network described earlier. All the native data is assigned a native transcription, thus only updating statistics of the native models during the Baum-Welch HMM training algorithm. The nonnative data is assigned a nonnative transcription, thus updating statistics of both the native and nonnative phone models. The assumption here is that some of the nonnative speakers will produce speech as a mixture of native and nonnative phones. The Baum-Welch algorithm will automatically assign the appropriate weight to each alternate path in the MP network during training.

To detect mispronounced phones we assume knowledge of the orthographic transcription. An MP network is assembled for the whole utterance by concatenating word models. The expanded network is searched using the Viterbi algorithm. The exact path is obtained from the Viterbi phone backtrace, which contains a sequence of native and nonnative phones. To evaluate overall pronunciation quality, we count the number of native and nonnative phones found in the backtrace as described in the next section.

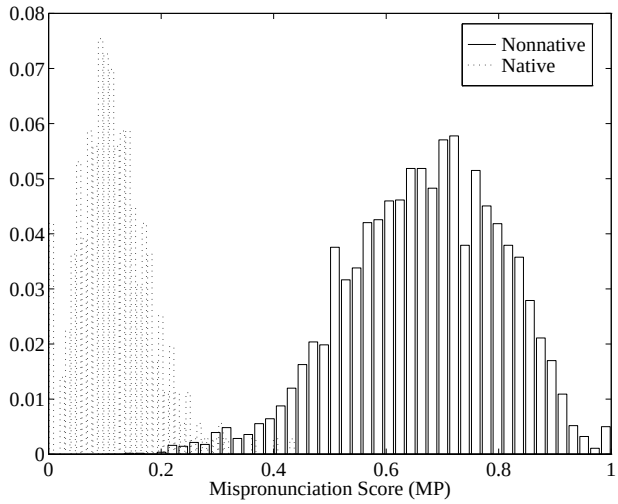


Figure 2. Normalized histogram of sentence-level mispronunciation scores for native and nonnative test utterances.

3. SCORING MISPRONUNCIATION

We developed techniques for scoring mispronunciation based on the new model described above. Given this model with two alternative pronunciations for each phone, we generate phonetic alignments of the student's speech. The result is the sequence of phones uttered by the student with a distinction between native and nonnative versions of each phone. From the phonetic alignment of each sentence, we compute a mispronunciation score (MP) that is the relative ratio of the number of nonnative phones to the total number of phones in the sentence:

$$\mathrm{MP} = \frac{\#(\text{nonnative phones})}{\#(\text{phones})}$$

A different approach for scoring mispronunciation with the new acoustic model is to perform forced alignments with two linear pronunciation networks (as in Figure 1.a): one consisting of only native phones and the other one consisting of only nonnative phones. We can use the likelihood scores from these forced alignments and compute a likelihood ratio score for a sentence, or a combination of these scores. The motivation for computing a likelihood ratio score is that it is a way to normalize the likelihood of the native models by the likelihood of the nonnative models.

To evaluate the performance of the model, we compute the correlation between the machine scores, such as the MP score and the likelihood score combination, and the human scores at the sentence and speaker levels.

4. EXPERIMENTAL RESULTS

For simplicity, we evaluate the mispronunciation model using context-independent (CI) phones. In previous work we used context-dependent (CD)

Model	Baseline		MP
	CI	CD	CI
Sentence	0.47	0.53	0.50
Speaker	0.82	0.85	0.84

Table 1. Human-machine correlation of the average phone posterior probability score. The baseline results are for CI and CD native models trained with a native pronunciation network. The MP results are for CI native models trained with the mispronunciation network shown in Figure 1.b.

phone models to compute phone-posterior-based pronunciation scores. To verify the effectiveness of the CI models, we first evaluate performance in pronunciation scoring when using CI and CD models. We trained the CI models using a normal linear network of native phones and using the proposed MP network. The results are shown in Table 1. We notice that the correlation with CI models is very close to that of the CD models, and that the MP native models are almost the same as the previous native models.

We performed forced alignments with the MP model and computed the MP score for the utterances in a nonnative test set and a native test set. Figure 2 shows a normalized histogram of sentence-level MP scores of native and nonnative test utterances. We can see that the average MP score for nonnative data is about 0.7, while for native data it is around 0.1. We also see a small region of overlap in scores between the two test sets (between 0.2 and 0.45).

When scoring with the proposed MP model, we found that some phones are less relevant than others. We weighted some phones less heavily than others, depending on their importance for mispronunciation. We gave the highest weight (=2.0) to the phones identified by teachers as the most problematic for American speakers. This subset of phones is the same set used in the work on scoring specific phone segments [8]. An intermediate weight (=0.5) was given to the other vowels and sonorants (semivowels, and nasal sounds). Finally, the lowest weight (=0.1) was given to the obstruents (the fricatives and stops). The weights were adjusted by experiments of measuring the correlation with different weight combinations. In Table 2 we show the human-machine correlation using the MP machine score. The correlation is negative because the human ratings measure the goodness of the pronunciation and the MP scores measure the mispronunciation level. The weighted MP machine scores weighs the occurrence of each phone according to its relevance. We can see that weighting the phones improves correlation at both sentence and speaker levels.

In the next experiment, we used the set of native models and the set of nonnative models of the MP model separately in testing. Each test utterance was decoded twice, using both model sets. For each set

Correlation	No Weight	Weight
Sentence	-0.29	-0.36
Speaker	-0.45	-0.58

Table 2. Human-machine correlation at the sentence and speaker levels for the mispronunciation score with and without weighting of the phone occurrences.

we computed the HMM log-likelihood scores. A likelihood ratio was computed by linearly combining the native and nonnative log-likelihood scores.

Table 3 shows the correlations with the weighted combination of the log-likelihood scores along with the correlations of the native and nonnative log-likelihood scores and of the native average phone posterior score. From the results we see that this normalization is very effective and it increases the correlation relative to the correlation of the individual log-likelihood scores. The correlation of the combined likelihood score is close to the correlation of the native posterior score. The weights for the score combination were optimized to maximize the correlation over a separate data set from the data set used to compute the correlation shown in the table. In estimating the weights for the linear combination of native and nonnative log-likelihood scores, we found that the weight of the native score was always positive while the weight of the nonnative score was always negative, and they both had the same order of magnitude. Since the scores are in a log-scale, this means that the score combination is a normalization of the native score by the nonnative score. The ratio between the magnitude of the nonnative and the native weights was 0.78 at the sentence-level correlation and 0.75 at the speaker-level correlation.

We performed an initial pilot experiment to validate the mispronunciation detection algorithm. We collected a small data set of 150 utterances from three native French speakers (2 males, 1 female). The data set contains five sentences read by all three speakers and repeated several times. The speakers were asked to mispronounce some of the phones in these sentences. The set of mispronounced phones were the set of 10 phones identified by teachers as the most problematic for American speakers. The experiment was conducted as follows. First, the speaker read the sentences normally (native version). Then, we marked four phones within each sentence and asked the speaker to repeat the sentence four more times and in each repetition to mispronounce one of the phones (nonnative version) while reading the rest of the sentence normally. Hence, we got 20 occurrences of mispronounced phones from each speaker, and we asked the speakers to repeat this twice, so that we have a total of 40 mispronunciations from each speaker. Since the speakers were asked to mispronounce one phone at a time, we also have 160 occur-

Correlation	Native Posterior	Native Likelihood	Nonnative Likelihood	Combined Likelihood
Sentence	0.50	0.29	0.06	0.44
Speaker	0.84	0.43	0.08	0.72

Table 3. Human-machine correlation at the sentence and speaker levels for the native phone posterior score, the native and nonnative log-likelihood scores, and a weighted combination of the two log-likelihood scores.

	Nonnative	Native
Occurrences	120	480
Error	24%	15%

Table 4. Misdetection error rate on the nonnative test set and false detection error rate on the native test set. The number of occurrences is the number of phone segments in each test set.

rences of native pronunciations of the same subset of phones for each speaker.

We evaluated the performance of the mispronunciation detection algorithm using this data by generating phonetic alignments and computing the detection rate on the nonnative and native test sets. The results are shown in Table 4. We asked a human expert to listen to the data and verify that in the nonnative set the phones were indeed mispronounced and in the native set they were pronounced correctly. It turned out that some of the machine errors (misdetection on the nonnative set and false detection on the native set) are in agreement with the human judgment. For example, in the nonnative set, in some cases, the mispronounced phones sounded like native phones to the human expert and to the machine. In other cases, the phones were mispronounced by changing the phone to another phone and inserting a phone. These variations were not detected by the system and it may be because the mispronunciation did not sound closer to nonnative than to native pronunciation. In the native set, one speaker was speaking fast in some parts of the sentences and, therefore, some phones in these regions were detected as mispronounced due to reduction. Another speaker may have a dialect influence on the pronunciation of a few phones that can cause some native phones to sound more nonnative. Based on these observations, we believe that the error rates shown in Table 4 are an upper bound for the errors we can get on a hand-labeled database of mispronunciations.

5. SUMMARY

We introduced a method for modeling mispronunciation. We evaluated the performance of the model by computing the correlation between the human and machine scores derived from the mispronunciation model.

We are continuing to work on this problem, and our goal is to use the mispronunciation model in a language instruction system to detect mispronunciation and to provide the student with precise feedback about pronunciation mistakes. We are currently gathering human ratings targeted at detecting mispronunciations of nonnative speakers, which we will use for direct evaluation of the mispronunciation model along the lines of the pilot study we performed.

6. ACKNOWLEDGMENT

This work was funded by the U.S. government under the TRP program and by internal sources. The views expressed here do not necessarily reflect those of the government.

REFERENCES

[1] L. Neumeyer, H. Franco, M. Weintraub, and P. Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Proceedings of ICSLP*, pages 1457–1460, Philadelphia, PA, October 1996.

[2] M. Rypa. VILTS: The voice interactive language training system. In *Proceedings of CALICO*, Albuquerque, New Mexico, 1996.

[3] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub. Automatic evaluation and training in English pronunciation. In *Proceedings of ICSLP*, Kobe, Japan, 1990.

[4] J. Bernstein. Automatic grading of English spoken by Japanese students. *SRI International Internal Reports,* Project 2417, 1992.

[5] V. Digalakis. Algorithm development in the Autograder project. *SRI International Internal Communication*, 1992.

[6] V. Digalakis and H. Murveit. GENONES: Optimizing the degree of mixture tying in a large vocabulary hidden Markov model based speech recognizer. In *Proceedings of ICASSP*, pages 537–540, 1994.

[7] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen. Automatic pronunciation scoring for language instruction. In *Proceedings of ICASSP*, pages 1471–1474, Munich, Germany, April 1997.

[8] Y. Kim, H. Franco, and L. Neumeyer. Automatic pronunciation scoring of specific phone segments for language instruction. In *Proceedings of EUROSPEECH*, Rhodes, Greece, September 1997.