AUTOMATIC PRONUNCIATION SCORING FOR LANGUAGE INSTRUCTION

Horacio Franco, Leonardo Neumeyer, Yoon Kim and Orith Ronen

Speech Technology and Research Laboratory
SRI International
http://www.speech.sri.com

ABSTRACT

This work is part of an effort aimed at developing computer-based systems for language instruction; we address the task of grading the pronunciation quality of the speech of a student of a foreign language. The automatic grading system uses SRI's Decipher™ continuous speech recognition system to generate phonetic segmentations. Based on these segmentations and probabilistic models we produce pronunciation scores for individual or groups of sentences. Scores obtained from expert human listeners are used as the reference to evaluate the different machine scores and to provide targets when training some of the algorithms. In previous work [1] we had found that duration-based scores outperformed HMM log-likelihood-based scores. In this paper we show that we can significantly improve HMM-based scores by using average phone segment posterior probabilities. Correlation between machine and human scores went up from r=0.50 with likelihood-based scores to r=0.88 with posterior-based scores. The new measures also outperformed duration-based scores in their ability to produce reliable scores from only a few sentences.

1. INTRODUCTION

The possibility of accepting speech input in computer-based language instruction systems allows developers to complement reading and listening comprehension with activities of production and conversation. In these systems, the computer may provide some feedback of the kind that an instructor would produce, such as an assessment of the quality of pronunciation or pointing to specific production problems or mistakes. Speech recognition technology is the key allowing such feedback. However, standard speech recognition algorithms were not designed with the goal of speech quality assessment; therefore, new methods and algorithms must be devised to approximate the perceptual capabilities of human listeners to grade speech quality.

The aim of this work is to develop methods for automatic assessment of pronunciation quality, to be used as part of a computer-aided language instruction system [1][2]. The basic pronunciation scoring paradigm [3][4][5] uses hidden Markov models (HMMs) [6] to generate phonetic segmentations of the student's speech. From these segmentations, we use the HMMs to obtain spectral match and duration scores. The effectiveness of the different machine scores is evaluated based on their correlation with expert human scores on a large database. Previous approaches were based on statistical models built for specific sentences [5]. The current algorithms were designed to produce pronunciation scores for arbitrary sentences, that is, sentences for which there is no acoustic training data [1]. This approach allows great flexibility in the design of language instruction systems because new pronunciation exercises can be added without retraining the scoring system.

We extend previous work [1] by introducing a new HMM-based score based on phone posterior probabilities. The level of human-machine correlation for this new score was significantly better than both likelihood and duration scores for the case of sentence specific scoring. When averaging scores across several sentences corresponding to a given speaker to obtain speaker-level scores we found that the new method required fewer sentences to achieve a similar level of correlation. We also investigated the combination of different machine scores to obtain a higher level of correlation. We experimented with linear and nonlinear regression as well as with an estimation-based approach to predict human scores from machine scores.

2. THE DATABASE

The requirements of data needed for development of the scoring system are more demanding than those typical of speech recognition systems [1]. A database of transcribed native read speech is used for training models for speech recognition and pronunciation scoring. A database of nonnative read speech is transcribed and scored for pronunciation quality at different levels of detail by expert human raters.

Speech was recorded from 100 natives of Parisian French (native corpus) and from 100 American students speaking in French (nonnative corpus). All the speech was recorded in quiet offices using a high-quality Sennheiser microphone.

A panel of five French teachers, certified language testers, rated the overall pronunciation of each nonnative sentence on a scale of 1 to 5 ranging from unintelligible to native quality. There was some overlap in the speech material rated by the teachers for consistency checking. The five teachers were selected from a group of ten based on their consistency in a pilot study. All the teachers were government-certified in language testing.

3. PRONUNCIATION SCORING

3.1. Human Scoring

The human scores are the reference against which the performance of the automatic scoring systems should be tested and calibrated; as such, it is important to assess the consistency of these scores both between raters (inter-rater correlations) and individually within each rater (intra-rater correlations).

Human judgments were provided by five raters on speech data from 100 students. A more detailed analysis of the human scores was presented in [1]. Two types of correlation were computed: at the *sentence level* pairs of corresponding ratings for all the individual sentences were correlated; at the *speaker level*, first, the scores for all the sentences from each speaker were averaged, and then the sequence of pairs of corresponding average scores for each of the speakers was correlated.

The consistency within and across raters was assessed in a subset of the database that was rated by all five raters and twice for each rater. The average sentence/speaker level inter-rater correlation was r=0.65/0.8; the average correlation between a rater and the average of a pool of the other raters was r=0.76/0.87. The average intra-correlation at the sentence level was r=0.76. These values may be considered upper bounds on what could be reasonably expected performance for the machine scoring system.

3.2. Automatic Scoring

The different pronunciation scoring algorithms studied are all based on phonetic time alignments generated using SRI's Decipher™ HMM-based speech recognition system [6]; these HMMs have been trained using the database of native speakers.

To generate the alignments for the student's speech we must know the text read by the student. We do this by eliciting speech in a constrained way in the language learning activities, and then backtracking the time-aligned phone sequence using the Viterbi algorithm. From these alignments, and statistical models obtained from the native speech, different probabilistic scores are derived for the student's speech. The statistical models used to do the scoring are all based on phone units and as such, no statistics of specific sentences or words are used. Consequently, the algorithms are text independent.

Here, we review some of the previously introduced scoring algorithms in [1] along with the newly introduced posterior probability-based score.

3.2.1. HMM-based phone log-likelihood scores

In this approach we use the HMM log-likelihood to derive a score. The underlying assumption is that the logarithm of the likelihood of the speech data, computed by the Viterbi algorithm, using the HMMs obtained from native speakers is a good measure of the similarity (or match) between the native speech and the students's speech. For each sentence the phone segmentation is obtained along with the corresponding log-likelihood for each segment. Then, for each phone segment we define the normalized log-likelihood $\hat{l}_i$ as

$$\hat{l}_i = l_i / d_i \,, \tag{1}$$

where l_i is the log-likelihood corresponding to the i -th phone and d_i is its duration in frames.

The likelihood-based score for a whole sentence L , is defined as the average of the individual normalized log-likelihood scores for each phone segment,

$$L = \frac{1}{N} \sum_{i=1}^{N} \hat{l}_i \tag{2}$$

where the sum run over the number of phones in the sentence N .

3.2.2. HMM-based phone log-posterior probability scores.

In this case we use a set of context-independent models along with the HMM phone alignment to compute an average posterior probability for each phone. First, for each frame belonging to a segment corresponding to the phone q_i we compute the frame-based posterior probability $P(q_i|y_t)$, of the phone i given the observation vector y_t :

$$P(q_i|y_t) = \frac{p(y_t|q_i)P(q_i)}{\sum_{j=1}^{M} p(y_t|q_j)P(q_j)} \tag{3}$$

where $p(y_t|q_i)$ is the probability density of the current observation using the model corresponding to the q_i phone. The sum over j runs over a set of context-independent models for all phone classes. $P(q_i)$ represents the prior probability of the phone class q_i .

Similarly to the previous case, the average of the logarithm of the frame-based phone posterior probability over all the frames of the segment is defined as the posterior score $\hat{\rho}_i$ for the i-th phone segment:

$$\hat{\rho}_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log P(q_i|y_t) \tag{4}$$

The posterior-based score for a whole sentence ρ is defined as the average of the individual posterior scores over the N phone segments in a sentence:

$$\rho = \frac{1}{N} \sum_{i=1}^{N} \hat{\rho}_i \,. \tag{5}$$

We expect that the posterior-based score could be less affected by changes in the spectral match due to particular speaker characteristics or acoustic channel variations. The same changes in acoustic match would affect both numerator and denominator similarly in Eq. (3), making the score more invariant to those changes and more focussed on the phonetic quality.

3.2.3. Segment duration scores

The procedure to compute the phone-based duration score is as follows: first, from the Viterbi alignment we measure the duration in frames for the i-th segment; then its value is normalized to compensate for rate of speech. To obtain the corresponding phone segment duration score, the log-probability of the normalized duration is computed using a discrete distribution of durations for the corresponding phone. The discrete duration distributions have been previously trained from alignments generated for the native training data. Again, the corresponding sentence duration score is defined as the average of the phone segment scores over the sentence

3.3. Combination of Scores

The combination of several different machine scores may allow to get a better prediction of the desired estimate of the human score. We investigated the use of linear and nonlinear regression as well as an estimation method.

In the linear regression we linearly combine two or more machine scores for each sentence, plus a bias term, to approximate the actual human score. The linear coefficients are optimized to minimize the mean square error between the predicted and the actual human scores over the sentences of the development set.

For the nonlinear regression the machine scores to be combined are the input to a neural network that computes the mapping between the multiple machine scores and the corresponding human scores. The actual human scores provide the targets for the training of the network. The network has a single linear output unit and 16 sigmoidal hidden units. It was trained with backpropagation using cross-validation on 15% of the training data. The training is stopped when performance degrades on the cross-validation set.

The regression approaches assume that the inputs are noiseless and that the only source of randomness is in the predicted variable; this assumption is clearly wrong in our case where both machine and human scores are highly noisy. To overcome this assumption we used an estimation procedure to get a prediction of the human scores based on the machine scores.

In this method the predicted human score $\tilde{h}$ is computed as the conditional expected value of the actual human score h given the measured machine scores $m_1, m_2, \ldots, m_n$:

$$\tilde{h} = E[h|m_1, m_2, \ldots, m_n]$$

To compute the expectation we need the conditional probability $P(h|m_1, m_2, \ldots, m_n)$ that we compute as

$$P(h|m_1, m_2, \ldots, m_n) = \frac{P(m_1, m_2, \ldots, m_n|h)P(h)}{\sum_{i=1}^{5} P(m_1, m_2, \ldots, m_n|h_i)P(h_i)}$$

where $P(h)$ is the prior probability of the score and the conditional distribution $P(m_1, m_2, \ldots, m_n|h)$ is modeled approximately by a discrete distribution based on scalar or vector quantization of the machine scores.

3.4. Experimental Results

We evaluated first, in terms of its level of correlation, the performance of the individual scores we have presented. Then, the effect of the number of sentences whose scores are averaged in the computation of correlation at the speaker level was studied. Finally, we evaluated the methods to combine the different types of machine scores in order to obtain a better prediction of the human scores.

3.4.1. Human-machine correlation of individual scores

We evaluated each of the proposed methods experimentally by computing the correlations between machine and human scores at the sentence and the speaker level. The speech material consisted of 5089 different sentences read from newspapers by 100 nonnative speakers. These sentences were rated at least once by one human rater. The different machine scores for each individual sentence were correlated with the corresponding human ratings.

When obtaining the machine scores for each sentence, in all the experiments we removed the scores of the phones in context with silence because their alignments may be inaccurate. Doing so produced a small but consistent increase in the correlation for all the machine score types.

At the speaker level, about 50 sentence scores were averaged for each of the 100 speakers before the correlation was computed. The results are presented in Table 1.

Algorithm	Correlation Coeff.	
	Sent. Level	Spkr. Level
Likelihood score	0.33	0.50
Posterior score	0.58	0.88
Normalized duration score	0.47	0.84

Table 1: Sentence- and speaker-level correlations between human and machine scores using 100 nonnative speakers and about 50 utterances per speaker.

We see that at the sentence level the posterior-based score has the highest correlation, followed by the duration score having a 20% lower correlation. At the speaker level the normalized duration and the log-posterior scores are comparable, rendering a performance similar to that of the human raters as we showed in Section 3.1. The log-likelihood score is the worst at both the sentence and speaker levels.

Sentence-level correlations are still lower than those among humans, suggesting that further work is needed to predict pronunciation ratings using only a single utterance.

3.4.2. Effect of different amounts of speaker data

We calculated the speaker-level correlation between human and machine scores using various amounts of test data. We varied the number of sentences per speaker (N) from 1 to 50 in obtaining the averaged score for each speaker. The human scores were the speaker averaged scores of the 100 speakers, using the entire human score data. We randomly chose N sentences per speaker to obtain the speaker-average machine score. We repeated this random experiment 40 times and averaged the correlation values for each N.

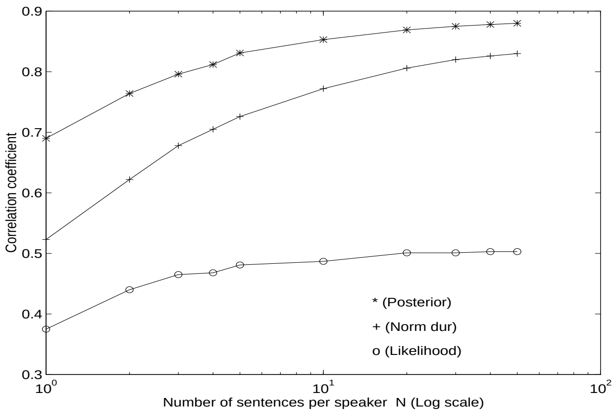

Fig. 1. Speaker level correlation for posterior, duration, and likelihood scores for different numbers of sentences per speaker

As we can see in Fig. 1, the posterior probability score performs the best for every N, but particularly well for low values of N, which is attractive for this application.

3.4.3. Combination of scores

We evaluated combining different types of machine scores in order to increase the correlation at the sentence level. We divided the 5089 sentences into two equally sized sets with no common speakers. We estimated the parameters of the regression and estimation models in one set and evaluated the correlation of the predicted and actual human scores in the other set. Then we repeated the procedure with the sets swapped and averaged the correlation coefficients.

Combination	Machine scores	correlation
-	posterior	0.58
Linear	posterior + duration	0.59
Nonlinear	posterior + duration	0.62
Estimation	posterior + duration (scalar quantization)	0.60
Estimation	posterior + duration (VQ)	0.62

Table 2: Sentence level correlations between human and combined machine scores.

In Table 2 we show the average correlation coefficients for the different types of score combination. Linear combination of posterior and duration scores produced a minor increase in correlation. The nonlinear combination using a neural network was more effective, increasing the correlation 7% with respect to that of the single posterior score. The estimation method using vector quantization of the scores was better than the one using scalar quantization and was comparable to the nonlinear combination method.

4. SUMMARY

We introduced a new HMM-derived score based on posterior probabilities of phone segments, and compared its performance with previously proposed pronunciation scores applied at both sentence and speaker levels.

At the sentence level, the posterior probability score had a 20% higher correlation with human scores than that obtained using duration scores. At the speaker level it also showed better performance, particularly when using few sentences to compute speaker-level scores.

An additional 7% increase in correlation at the sentence level was obtained by combining posterior and duration scores using nonlinear regression with a neural network, or, alternatively, using an estimation method to predict the human scores given the machine scores.

5. ACKNOWLEDGEMENT

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.

6. REFERENCES

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

2. M. Rypa, “ECHOS: A Voice Interactive Language Training System”, Proceedings of *CALICO*, Alburquerque, New Mexico, 1996.

3. J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, “Automatic Evaluation and Training in English Pronunciation”, *ICSLP 1990, Kobe, Japan.*

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,” Proc. of *ICASSP94*, pp. I537-I540, 1994.