Combination of Machine Scores for Automatic Grading of Pronunciation Quality

Horacio Franco, Leonardo Neumeyer, Vassilios Digalakis, and Orith Ronen

Speech Technology and Research Laboratory
SRI International

February 28, 1998

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 different pronunciation scores for individual or groups of sentences that can be used as predictors of the pronunciation quality. Different types of these machine scores can be combined to obtain a better prediction of the overall pronunciation quality. In this paper we review some of the best-performing machine scores, and discuss the application of several methods based on linear and nonlinear mapping and combination of individual machine scores to predict the pronunciation quality grade that a human expert would have given. We evaluate these methods in a database that consists of pronunciation-quality-graded speech from American students speaking French. With predictors based on spectral match and on durational characteristics, we find that the combination of scores improved the prediction of the human grades and that nonlinear mapping and combination methods performed better than linear ones. Characteristics of the different nonlinear methods studied are discussed.

Keywords: Automatic pronunciation scoring; Combination of scores; Hidden Markov models; Speech Recognition; Pronunciation quality assessment; Language instruction systems; Computer aided language learning.

1 Introduction

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]. Typical foreign language instruction courses focus mainly on reading, writing, and listening comprehension, much less effort is devoted to teaching correct pronunciation. One of the reasons possibly being that it requires more expensive resources, such as extensive practice with private tutors who are natives of the target language. An interactive system capable of grading the pronunciation quality could facilitate the pronunciation learning process by giving feedback on the students ability and progress to produce the foreign language sounds. A prototype application [3] allows students to listen to natives saying phrases, or selected parts of phrases, and record themselves to obtain pronunciations scores. Segments that are difficult to produce can be practiced by selecting the target words and obtaining scores for them. The content can be easily updated to provide new lessons with minimum effort.

The basic pronunciation scoring paradigm [4][5][6] uses hidden Markov models (HMMs) [7] 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 human grades on a large database. Previous approaches were based on statistical models built for specific sentences [6]. 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.

In this work we focus on the problem of the mapping and combination of different machine scores to obtain a better prediction of the human grades. We experimented with linear and nonlinear regression based on neural networks and regression trees as well as with an estimation-based approach to predict human grades 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. Here we describe part of the data collected under the VILTS (Voice Interactive Language Training System) project [1]. One database of transcribed native speech is used for training models for speech recognition and pronunciation scoring. A second database of nonnative read

speech is transcribed and graded for pronunciation quality at different levels of detail by expert human raters; it is used to develop and calibrate the pronunciation scoring algorithms.

The native speech database consisted of 16,000 sentences recorded from 100 native speakers of Parisian French. The recordings consisted of different read newspaper sentences, with no common sentences across speakers to maximize the coverage of different words and contexts. The average length of a sentence was 19 words.

The nonnative corpus used for this study consisted of 5089 different sentences read from newspapers by 100 American students speaking in French. The average length of the nonnative sentences was 14 words. We divided the 5089 sentences into two equally sized sets, with no common speakers, that were used alternatively as development and evaluation sets. All the speech was recorded in quiet offices by a high-quality Sennheiser microphone. The overall pronunciation of each of the 5089 nonnative sentences was rated on a scale of 1 to 5 by human experts. There was some overlap in the speech material rated by the teachers for consistency checking.

3 Pronunciation Scoring

3.1 Human Scoring

The human grades 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 grades both between raters (inter-rater correlations) and within each rater (intra-rater correlations) when multiple ratings for the same sentence are given. Two types of correlation were computed. At the *sentence level,* pairs of corresponding ratings for any pair of raters for all the individual sentences were correlated. At the *speaker level*, first, the grades for all the sentences from each speaker were averaged, and then the sequence of pairs of corresponding average grades for each of the speakers was correlated.

In a preliminary study, five French teachers, certified language testers, were selected from a group of ten candidates as the five most self-consistent raters. This panel of five teachers rated the overall pronunciation quality of each of the 5089 nonnative sentences on a scale of 1 to 5, ranging from the categories “strongly nonnative” to “almost native”. The probability distribution of grades obtained is shown in Table 1 [1], where we appreciate that the distribution peaks around grades 2 and 3. The consistency across raters was assessed in a subset of 342 nonnative sentences that were rated by all five raters. 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 consistency within raters was assessed in different subsets of approximately 130 sentences per rater that were rated twice by each rater. The average intra-correlation at the sentence level was r=0.77. These values may be considered upper bounds on the expected performance for the machine scoring system. The values of the correlation coeffi cients computed in this work are dependent on the probability distribution of grades given above; as such, the correlation values should be used as relative measures.

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 [7]; these HMMs have been trained using the database of native speakers. The front-end extracts mel-frequency cepstral coeffi cients (MFCC); the MFCC have the mean over each sentence removed for acoustic channel normalization purposes. Both, context-independent and triphone (context-dependent) models were trained. For recognition and forced alignments we used the context-dependent models, while the context-independent models were used for one of the scoring measures. The recognition system used Gaussian mixture models for computation of the observation probabilities. 100 mixture components were used in each Gaussian mixture. Also, the Gaussians components were shared across all the triphone models corresponding to a given center phone, as well as its corresponding context-independent model. This type of acoustic model is usually referred to as " phonetically tied mixtures". The conte xt-independent phone classes corresponded to the standard set of phonemes in French with one borrowed consonant added (/ng/).

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 by using the Viterbi algorithm. From these alignments, and statistical models obtained from the native speech, 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 specifi c sentences or words are used. Consequently, the algorithms are text independent. Here, we review some of the scoring algorithms introduced in [1] and [8].

3.2.1 HMM-based phone log-posterior probability scores

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:

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

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 the set of all context-independent phone models. $P(q_i)$ represents the prior probability of the phone q_i.

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:

$$\hat{\rho}_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log P(q_i|y_t) \quad (2)$$

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:

$$\rho = \frac{1}{N} \sum_{i=1}^{N} \hat{\rho}_i \,. \quad (3)$$

The log-posterior score is robust against changes in the spectral match that are due to particular speaker characteristics or to acoustic channel variations. This is a desirable property of pronunciation-evaluation scores, and can be attributed to the fact that the same changes in acoustic match affect both numerator and denominator similarly in Eq. (1), making the score fairly invariant to those changes and focused on the phonetic quality.

3.2.2 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 context-independent 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. Based on previous research [1], duration measurements were normalized by rate of speech (ROS) prior to the computation of the duration score. Therefore, the duration score is defined as

$$D = \frac{1}{N}\sum_{i=1}^{N} \log[p(f(d_i)|q_i)] \tag{4}$$

where d_i is the duration of the i-th segment corresponding to phone q_i, and $f(d_i) = d_i \cdot ROS$ is the duration normalization function. The ROS was estimated for each individual sentence as the average number of phones per unit of time. Segments labeled as silence as well as phones in context with silence were excluded from the ROS computation based on previous research [1].

3.2.3 Timing Scores

Given that language learners tend to impose the rhythm of their native language on the language they are learning, some measure of timing could represent the degree of fluency of a student and be used as a predictor of pronunciation quality. For example, English is stress-timed (stressed syllables tend to be lengthened and others shortened), while Spanish and French are syllable-timed. To attempt to exploit this aspect of linguistic timing, a distribution of normalized syllabic periods is estimated from natives and used to evaluate timing scores for nonnative speakers. The syllabic period is defined as the time interval between the centers of vowels within segments of speech. The normalization is based, as before, on sentence ROS. The procedure to obtain the timing distribution is similar to that used for the duration scores, a single discrete distribution is trained based on ROS-normalized syllabic periods in the force-aligned native database. Similarly to the duration score, the timing score is computed by taking the average of the log-likelihoods of the normalized syllabic periods over a given sentence.

3.3 Combination of Scores

The grade that an ideal, experienced human rater would assign to an utterance when rating either the general pronunciation quality or a particular skill can be treated as a random variable. The pronunciation evaluation problem can then be defined as an estimation problem, where we try to estimate, or predict, the value of the ideal human grade h by using a set of predictors. These predictors are the machine scores that we have presented in our previous work [1]-[7], some of which were summarized in the previous section.

We investigated the use of linear and nonlinear regression methods to predict one random variable, the human grade, from a set of others, the machine scores.

Applying a well-known result from probability theory [10], when using a minimum mean square error criterion (Eq. 5) between the actual human grades h and the predicted ones $\tilde{h} = d(m_1, m_2, \ldots, m_n)$

$$\min_{d} \quad E[h - d(m_1, m_2, \ldots, m_n)]^2 \tag{5}$$

the optimal predictor of the human grade $\tilde{h}_{opt}$ is the conditional expected value of the actual human grade h given the measured machine scores $m_1, m_2, \ldots, m_n$, that is

$$\tilde{h}_{opt} = E[h|m_1, m_2, \ldots, m_n] \quad . \tag{6}$$

3.3.1 Linear regression

If the machine and human scores can be modeled as jointly Gaussian random variables or there is a linear relationship among them, then the conditional expected value of the human grade is a linear combination of two or more machine scores for each sentence plus a bias term. Thus, the predicted human grade $\tilde{h}$ is

$$\tilde{h} = a_1 m_1 + a_2 m_2 + \ldots + a_n m_n + b \ . \tag{7}$$

The linear coeffi cients $a_1, \ldots, a_n, b$ are optimized [9] to minimize the mean square error between the predicted and the actual human grades over the sentences in a development set. Linear regression is a parametric approach; as such, the number of parameters that have to be estimated is reasonably small, which leads to simple and robust estimates.

3.3.2 Nonlinear regression

In the general case, this estimator is a nonlinear function of the machine scores. If we do not know the mathematical form of the underlying joint probability distribution of the human and machine scores, it is necessary to resort to nonparametric methods. The potential advantage of the nonlinear nonparametric methods is that they could attain a better approximation to Eq. (6) than the linear estimate of Eq. (7), by not having to assume a particular form for the distributions and by being able to model nonlinear maps between machine scores and human grades as well as nonlinear relationships between the machine scores.
Neural Networks

A neural network can be considered as a very flexible function approximator capable of implementing arbitrary maps between input and output spaces [11]. Its parameters, the weights, are adjusted by the train-

ing algorithm to minimize the training criterion. With this approach, the machine scores to be combined are the input to a neural network that computes the mapping between the multiple machine scores $m_1, m_2, \ldots, m_n$ and the corresponding predicted human grade $\tilde{h}$, that is

$$\tilde{h} = o(m_1, m_2, \ldots, m_n) \tag{8}$$

where $o(\ )$ represents the nonlinear mapping implemented by the network. The actual human grades provide the targets --or desired output values-- for the training of the network.

If the neural network is trained using the minimum mean square error criterion (5), if it has enough number of weights and layers, and if the training does not get stuck in a local minimum, the output of the neural network $o(m_1, m_2, \ldots, m_n)$ will closely approximate Eq. (6), that is, the conditional expected value of the desired output given the inputs [12],[13]:

$$o(m_1, m_2, \ldots, m_n) \cong E[h|m_1, m_2, \ldots, m_n] \quad . \tag{9}$$

Training algorithms for neural networks, such as backpropagation [14], only assure the attainment of a local minimum of the criterion. In addition, to obtain good performance, in practice, we have to use a large number of parameters for the network and therefore we need to use a regularization method to avoid overfitting to the training data. Typically, early stopping based on performance on a cross-validation set is used. In this way, not even a local minimum of the training criterion could be attained. Nevertheless, cross-validation-based training allows us to robustly train large networks, which in turn results in a good approximation to the optimal solution of Eq. (6).

After some preliminary experimentation with different network architectures, we chose an architecture consisting of a two-layer perceptron with a single linear output unit and a hidden layer of sigmoidal units. We varied the number of hidden units from 8 to 32; the best performance was obtained with 16 hidden units. The number of input units corresponded to the number of machine scores combined. The network was trained with backpropagation using the mean square error criterion. A momentum term was used in the weight update rule [14]. To avoid overfitting to the training data and to obtain good generalization, we used a cross-validation set formed with 15% of the training data. Prediction performance was assessed after each training iteration on this set; the training was stopped when performance, measured by the same error criterion, did not improve on the cross-validation set.

3.3.2.1 Probability distribution estimation

In this approach we approximate the conditional expectation (Eq. 6) by direct computation of the expected value by using estimates of the necessary conditional probabilities. The predicted human grade $\tilde{h}$ is computed as

$$\tilde{h} = E[h|m_1, m_2, \ldots, m_n] = \sum_{i=1}^{G} h_i \cdot P(h_i|m_1, m_2, \ldots, m_n) \tag{10}$$

where $P(h_i|m_1, m_2, \ldots, m_n)$ is the estimated conditional probability of the human grade h_i given the machine scores and G is the number of distinct grades. Taking advantage of the fact that the human grades are discrete variables (in this case in the range of 1 to 5), by using Bayes rule we can express this probability as

$$P(h_i|m_1, m_2, \ldots, m_n) = \frac{P(m_1, m_2, \ldots, m_n|h)P(h_i)}{\sum_{j=1}^{G} P(m_1, m_2, \ldots, m_n|h_j)P(h_j)} \tag{11}$$

where $P(h_i)$ is the estimated prior probability of the human grade h_i and $P(m_1, m_2, \ldots, m_n|h_i)$ is the estimate of the conditional distribution of the machine scores for a given human grade h_i.

In this work we modeled $P(m_1, m_2, \ldots, m_n|h_i)$ by discrete distributions that were estimated based on the quantization of the machine scores.We studied both scalar and vector quantization (VQ) cases. In the scalar case, we linearly quantized each machine score and estimated the joint discrete distribution of the combination of quantized machine scores. In the VQ case, the joint distribution of machine scores was modeled by a single discrete distribution on the VQ index $V(m_1, \ldots, m_n)$.

$$P(m_1, m_2, \ldots, m_n|h) \cong P(V(m_1, m_2, \ldots, m_n)|h_i) \tag{12}$$

In the design of the scalar quantizer we experimented with a different number of bins for each machine score. We searched for the maximum correlation in the range of 5 to 20 bins when combining scores, and in the range of 5 to 100 bins when mapping posterior scores only. We found that when there are too few bins or too many bins, the correlation is low. We obtained the best result with 10 bins when combining three scores, 12 bins when combining two scores, and 50 bins when using a single machine score. Bin probabilities were estimated by frequency counts on the training data and smoothed by flooring their values. We designed the vector quantizer (VQ) with K codewords for the machine score combination, using

the Euclidean distance as the distortion measure; the individual machine scores had been previously scaled to zero mean and unit variance over the whole database. We initialized the VQ by choosing K random data points from the training set, and then we performed 15 iterations of the LBG algorithm. For the VQ case, we estimated the discrete distribution of the machine score combination for each human grade by using smoothed relative frequency. We tested codebooks of K=10 to K=100 codewords. The best results were $K = 20$ codewords for posteriors, $K = 15$ codewords for posterior and duration, and $K = 50$ codewords for posterior, duration, and timing.

The vector quantization approach, which is more efficient in the use of parameters, resulted in more accurate and robust estimates of the distributions, as was reflected in the experimental results.

3.3.2.2 Regression trees

In the previous section we estimated the conditional posterior distribution of the human grade, given the machine scores, by using Bayes rule and the class distributions of the quantized machine scores. An alternative approach is to estimate the conditional posterior distribution directly, using classification and regression trees [15]. A tree can be used to classify a vector of machine scores $(m_1, m_2, \ldots, m_n)$ to one of several possible classes $\{t_1, t_2, \ldots, t_N\}$, each class representing a final node (a leaf) of the tree. The conditional distribution of the human grade, given a set of machine scores, is then approximated by

$$p(h|m_1, m_2, \ldots, m_n) \cong p(h|t) \tag{13}$$

where t is the leaf corresponding to the machine scores $m_1, m_2, \ldots, m_n$.

Specifically starting at the root of the tree, a question is asked at each node, resulting in a choice of one of two branches leaving that node; the process is repeated until a leaf node is reached. Each leaf represents a subset of the training data with similar or homogeneous properties, and an estimate of the conditional distribution (13) can be obtained using this data. The estimate of the human grade assigned to that leaf is the mean of that conditional distribution, obtained by combining Eqs. (10) and (13):

$$\tilde{h}(t) = E(h|t) \cong \sum_{i=1}^{G} h_i \cdot p(h_i|t) \tag{14}$$

When an input machine score vector ends up in a leaf, the predicted human grade is the one assigned to that leaf. With trees, the machine score's input space is partitioned by the sequence of binary splits into regions assigned to leaf nodes. In each leaf node t, the predicted response value (the predicted human

grade) $\bar{h}(t)$ is constant; therefore, the tree can be considered as a histogram-like estimate of the regression surface.

By using the supervised training data composed of machine scores $(\overline{m}, h)$ and the corresponding human grades (where $\overline{m}$ represents the vector of machine scores $m_1, m_2, \ldots, m_n$), we can build the tree and obtain the rule that assigns the predicted human grade to each leaf with the aid of available tree construction algorithms [15]. The mean of the conditional distribution computed in Eq. (14) can also be computed as the sample mean of the human grades h corresponding to all $N(t)$ cases that fall into node t:

$$\tilde{h}(t) = \frac{1}{N(t)} \sum_{\overline{m} \in t} h_n \quad . \tag{15}$$

This predictor minimizes the mean square error over the training samples for a given tree structure. It is important then, to grow a tree structure in a way that would achieve the minimum of the global mean square error computed over all the data, which in turn would approximate the optimal estimator (Eq. 6). The procedure is to generate, at each node, the split that maximizes the decrease of the accumulated square error for the samples at the node being split.

In practice, we took a rather heuristic approach to growing the tree, as it was important to have both small size and good prediction. We used a public domain software package [16] to test different splitting criteria and pruning levels. The predicted scores were then correlated with the human grades. We chose the criterion that allowed us to obtain the best correlation between the predicted and the actual human grades in a development set. The heuristic splitting rule based on information gain [16][17] produced slightly better results. For each different number of input features, the depth of the trees was optimized by exhaustive search in the range of 3 to 10 levels; then, different degrees of pruning were explored. For each case, we chose the tree with the highest correlation between the predicted grades and the human grades; we picked the smallest trees in cases of equal performance. For the case that combines three machine scores, the best tree had a depth of 5 levels. A depth of 4 levels was optimal for the combination of two scores and for the scalar case. Best results were obtained with mild pruning which removed option branches whose proportion was less than 0.01.

3.4 Experimental Results

We evaluated each individual machine score in terms of its level of correlation with human grades. Then, we evaluated methods of combining the different types of machine scoring to obtain a better prediction of the human grades.

3.4.1 Human-machine correlation of individual scores

We evaluated each of the proposed scoring methods experimentally by computing the correlations between machine and human scores at the sentence level. The speech material consisted of the 5089 sentences of the nonnative database which were graded for pronunciation quality by expert human raters. The machine scores for each individual sentence were correlated with the corresponding human ratings. For the sentences that had more than one human grade assigned, we picked randomly one of them.

In all the experiments, when obtaining the machine scores for each sentence, we removed the scores of the phones in context with silence because their alignments could have been inaccurate. The results was a small but consistent increase in the correlation for all the machine score types [1].

In Table 2 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, and then the syllabic timing with 39% lower correlation. Sentence-level correlations are still lower than those among humans, which motivates further work to predict pronunciation ratings for only a single utterance.

We would consider attempting to increase human-machine correlation by using multiple scores. This approach is practical, however, only if the machine scores are not highly dependent on each other. As shown in Table 3, we computed the correlation coeffi cient between pairs of machine scores, and we found a moderate level of correlation between posterior and duration scores, and a lower degree of correlation between syllabic timing and both posterior and duration scores. Assuming that the joint probability distribution of the machine scores does not deviate very much from a multivariate Gaussian, and given that the correlations between the scores are not very high, we consider that there may be some independent information in the different machine scores, so that the combination could help to predict pronunciation quality.

3.4.2 Combination of scores

We evaluated the four different types of predictors --linear regression, neural network, probability distribution estimation, and regression tree-- in mapping and combining different types of machine scores to

increase the correlation at the sentence level. As referred in section Section 2, the nonnative speech database was divided into two equally sized sets with no common speakers. We estimated the parameters of the different regression and estimation models in one set, and we evaluated the correlation of the predicted scores and the corresponding actual human grades in the other set. Then we repeated the procedure with the sets swapped, and we averaged the correlation coefficients.

In Table 4 we show the average correlation coefficients for the diferent types of predictors and score combinations. A linear combination of posterior and duration scores produced a minor increase in correlation over the use of posterior scores alone. The addition of syllabic timing resulted in a small additional improvement.

The nonlinear combination using a neural network was more effective, in the best case increasing the correlation 11.5% with respect to that of the single posterior score. Table 4 shows that a big part of this gain, 8%, is due only to the effect of the nonlinear mapping of the posterior scores to the human grades that the network is implementing. The addition of the duration score led to a gain of 10.8% with respect to the baseline. The further addition of timing scores allowed to reach the best case, with a correlation of $r = 0.642$, which is very close to the sentence-level human-to-human correlation reported in Section 3.1. Assuming independent measurements, this improvement was statistically significant at the level of 0.95.

The distribution estimation method with scalar quantization produced a good prediction ($r = 0.611$) of the human grades when using only the posterior score, but degraded rapidly when the number of input features was augmented. To maintain a robust distribution estimate, we could make only a small increase in the number of parameters. Therefore, when adding a second machine score, we had to reduce the number of quantization levels. Thus, the quantization error increased, reducing the correlation.

The estimation method with vector quantization of the scores was better than the method with scalar quantization and gave, in the best case, a 7.3% improvement in correlation over the case with a single posterior score. This improvement was also statistically significant at the level of 0.95. Most of the gain was produced by the nonlinear mapping of the posterior scores rather that by the addition of more machine scores.

The tree approach achieved slightly better performance than the estimation method. It showed a marked increase in prediction performance with the addition of new input scores. The improvements ranged from 5.7%, with posterior scores only, to the best-performing case, with posterior, duration, and timing, which reached an 8% increase in correlation with respect to the baseline.

Our experiments showed that adding other machine scores, such as global likelihood and phone recognition rate [1], to the posterior score did not lead to an improvement.

Another consideration is that the correlation results we have shown are very dependent on the distribution of grades in the particular speech database. The VILTS database, for instance, has a concentration around the intermediate human grades of 3, 2, and 4 and fewer examples of 5 and 1 [1]. A more even distribution across all the grades would tend to have a higher human-machine correlation because having more examples of speech with the highest and lowest human grades exercises the ranges of machine scores that are more reliable or have less superposition with those associated to other human grades. For instance, when we increased the native sentences in the database by 7.5%, and assigned the human grade 5 to those sentences, the baseline correlation using posterior scores increased to $r = 0.64$ and the correlation between the human and the predicted grades, using a regression tree, rose to $r = 0.71$. This represents an increase in correlation of 10.7%, which is almost twice the increment obtained with the tree in the original database.

4 Discussion and Summary

We have presented and experimentally evaluated several linear and nonlinear methods of predicting human pronunciation-quality grades based on machine scores. Nonlinear predictors based on a combination of machine scores produced better results than a linear combination of scores. Each of the nonlinear combination methods allowed us to approximate in different ways the optimal predictor of the human grades, the conditional expected value given by Eq. 6. For the specific machine scores used in this work, the nonlinear mapping implemented by the nonlinear predictors was more effective than a combination of machine scores.

Each method presents different tradeoffs in terms of implementation difficulty, speed, and necessary degree of manual tuning and optimization. In the case of the neural network predictor we had to experiment with different network architectures to find the with the best performance. This method had a higher computational cost in training than other methods, and it required manual tuning of training parameters. In addition, it was not easy to interpret the processing of the neural network. The VQ-based estimation of distributions also required substantial experimentation to define the quantization levels or codebook size needed for good performance. The method based on regression with trees was very efficient in development time because the architecture of the tree is generated automatically, training is not computationally expensive, and humans can easily interpret the processing of the tree. In addition, it can deal easily with missing features, which is not true for the other methods. On the other hand, both trees and the VQ-based

prediction produce only a set of possible discrete values as the predicted human grades, while the neural network method produces a continuous variable, which may be a desirable feature in some applications, such as pronunciation quality "meters". The continuous-variable output combined with well-known interpolating capabilities may have contributed to neural network's performance, which was better than that of the other methods.

In the best case, an overall gain of 11.5% increase in correlation (from $r = 0.576$ to $r = 0.642$) at the sentence level was obtained in the VILTS database by using nonlinear regression with a neural network combining posterior, duration, and timing scores.

Acknowledgment

We gratefully acknowledge support from the U.S. Government under the Technology Reinvestment Program. 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," *Proc. of ICSLP 96*, pp. 1457-1460, Philadelphia, Pennsylvania, 1996.

[2] M. Rypa, "ECHOS: A Voice Interactive Language Training System," *Proceedings of CALICO*, Albuquerque, New Mexico, 1996.

[3] L. Neumeyer, H. Franco, V. Abrash, L. Julia, O. Ronen, H. Bratt, J. Bing, V. Digalakis, M. Rypa, "Web Grader: A Multilingual Pronunciation practice Tool", *Proceedings of Workshop on Speech Technology in Language Learning (STiLL) 98,* pp. 61-64, Marholmen, Sweden, 1998.

[4] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic Evaluation and Training in English Pronunciation," *ICSLP 1990*, pp. 1185-1188, Kobe, Japan.

[5] J. Bernstein, "Automatic Grading of English Spoken by Japanese Students," *SRI International Internal Reports Project 2417*, 1992.

[6] V. Digalakis, "Algorithm Development in the Autograder Project," *SRI International Internal Communication*, 1992.

[7] V. Digalakis and H. Murveit, "GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer," *Proc. of Proc. Intl. Conf. on Acoust., Speech and Signal Processing 94,* pp. 1537-1540, 1994.

[8] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic Pronunciation Scoring for Language Instruction," *Proc. Intl. Conf. on Acoust., Speech and Signal Processing 97,* pp. 1471-1474, Munich, 1997.

[9] N. Draper and H. Smith, "Applied Regression Analysis, 2nd Ed.," J. Wiley & Sons, 1981.

[10] S. Kay, "Fundamentals of Statistical Signal Processing," Apprentice Hall Signal Processing Series, A. Opening, Series Ed., 1993.

[11] Cybenko, G., "Approximation by Superpositions of a Sigmoidal Function." *Mathematics of Control, Signals, and Systems*, 2, pp. 303-314, 1989.

[12] H. Bourlard and C. Wellekens, "Links Between Markov Models and Multilayer Perceptrons," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 12, No. 12, pp. 1167-1178, 1990.

[13] M. Richard, R. Lippman, "Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities," *Neural Computation 3*, 461-483, 1991.

[14] D. Rumelhart and J. Mc Clelland, "Parallel Distributed Processing," MIT Press, 1986.

[15] L. Breiman, J. Friedman, R. Olsen, and C. Stone, "Classification and Regression Trees," The Wadsworks & Brooks/Cole Statistics/Probability Series, 1984.

[16] W. Buntine and R. Caruana, "Introduction to IND V2.1 and Recursive Partitioning," NASA Ames Research Center, 1992.

[17] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, 1 (1):81-106, 1986.

.

Scores	1	2	3	4	5
%	9	31	42	15	3

Table 1: Distribution of human grades over the nonnative database

.

Machine scores	sentence-level correlation
Posterior score	0.579
Normalized duration score	0.469
Syllabic timing score	0.352

Table 2: Sentence-level correlations between human and machine scores

Pairs of machine scores	sentence-level correlation
Posterior - duration	0.662
Posterior - syllabic timing	0.434
Duration - syllabic timing	0.466

Table 3: Sentence-level correlations between pairs of machine scores

Method	Machine scores	Correlation
(Baseline)	posterior	0.576
Linear regression	posterior + duration	0.587
Linear regression	posterior + duration + timing	0.593
Neural network	posterior	0.622
Neural network	posterior + duration	0.638
Neural network	posterior + duration + timing	0.642
Distribution estimation (scalar quantiz.)	posterior	0.611
Distribution estimation (scalar quantiz.)	posterior + duration	0.605
Distribution estimation (scalar quantiz.)	posterior + duration +timing	0.568
Distribution estimation (vector quantiz.)	posterior (VQ)	0.615
Distribution estimation (vector quantiz.)	posterior + duration (VQ)	0.617
Distribution estimation (vector quantiz.)	posterior + duration + timing (VQ)	0.618
Regression tree	posterior	0.609
Regression tree	posterior + duration	0.618
Regression tree	posterior + duration + timing	0.622

Table 4: Sentence-level correlations between human and predicted machine grades using different predictors and combinations of machine scores.