CALIBRATION OF MACHINE SCORES FOR PRONUNCIATION GRADING

Horacio Franco and Leonardo Neumeyer

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

ABSTRACT

Our proposed paradigm for automatic assessment of pronunciation quality uses hidden Markov models (HMMs) to generate phonetic segmentations of the student's speech. From these segmentations, we use the HMMs to obtain spectral match and duration scores. In this work we focus on the problem of calibrating different machine scores to obtain an accurate prediction of the grades that a human expert would assign to the pronunciation. We discuss the application of different approaches based on minimum mean square error (MMSE) estimation and Bayesian classification. We investigate the characteristics of the different mappings as well as the effects of the prior distribution of grades in the calibration database. We finally suggest a simple method to extrapolate mappings from one language to another.

1. INTRODUCTION

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 an HMM-based continuous speech recognition system [1] 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 estimation of the overall pronunciation quality. In this work we discuss the application of several methods to obtain and calibrate the mapping from the machine scores to the pronunciation quality grades that a human expert would have given. Treating these human grades and machine scores as random variables, the pronunciation evaluation problem can be considered as an estimation problem, where we try to estimate, or predict, the value of the human grade by using a set of predictors. These predictors are the machine scores that we have presented in our previous work [2],[3],[5].

We investigate the use of MMSE estimation and classification methods to predict the human grade from a set of machine scores. We present alternative implementations of these methods based on nonparametric techniques. We illustrate their application using a pronunciation-quality-graded database of nonnative Spanish. We also investigate the effect of the grade priors on the mappings. Finally, we suggest a simple method to extrapolate calibrated mappings from one language to another.

2. PRONUNCIATION SCORING

The different pronunciation scoring algorithms studied are all based on phonetic time alignments generated using SRI's Decipher™ HMM-based speech recognition system [1]; 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 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 specific sentences or words are used. Consequently, the algorithms are text independent. The computation of the scoring algorithms has been described in detail in [2] and [3]. We review only the most useful score here.

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.

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 for the *i*-th phone segment. The posterior-based score for a whole sentence is defined as the average of the individual posterior scores over the N phone segments in a sentence. The log-posterior score is fairly robust against changes in the spectral match due to particular speaker characteristics or acoustic channel variations.

3. CALIBRATION APPROACHES

Two approaches can be devised to obtain the mappings from machine scores to human pronunciation quality ratings, one based on MMSE estimation and other based on minimum error classification.

3.1. Estimation Approach

The grade a 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 the value of the ideal human grade h by using a set of predictors, the machine scores, that we have obtained from the speech sample to be graded.

Applying a well-known result from probability theory, when using a minimum mean square error criterion (Eq. 1) between the actual human grades h and the predicted ones $\tilde{h} = d(\bar{m})$,

$$\min_{d} \quad E[h - d(\bar{m})]^2, \tag{1}$$

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

$$\tilde{h}_{opt} = E[h|\bar{m}]. \tag{2}$$

In the general case this estimator is a nonlinear function of the machine scores.

3.2. Classification Approach

Taking the alternative approach, the mapping of machine scores to human grades can be casted as a classification problem. Each sentence is classified as belonging to one of N classes, where the classes are the discrete pronunciation grades assigned by the human raters. To classify a sentence we use the optimal Bayes' decision rule, which minimizes the misclassification error rate.

The optimal predictor of the human grade, $\tilde{h}_{opt}$, is the grade h_i that has the highest posterior probability given the machine scores $\bar{m}$,

$$\tilde{h}_{opt} = \arg\max_{i} [P(h_i|\bar{m})]. \tag{3}$$

Methods to compute Eq. (3) are similar to those used to compute Eq. (2), as both can be based in the computation of the posterior probabilities $P(h_i|\bar{m})$ given that Eq. (2) can be written as

$$E[h|\bar{m}] = \sum_{i=1}^{G} h_i \cdot P(h_i|\bar{m}). \tag{4}$$

For both the estimation and the classification approaches, 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. Useful nonparametric nonlinear methods to predict the human grades that we have investigated [4] are: neural networks, regression and classification trees and probability distribution estimation using scalar or vector quantization.

3.3. Nonparametric methods

Three possible implementations of the estimation and classification approaches using nonparametric methods are briefly reviewed.

Neural Networks. These are very flexible function approximators capable of implementing arbitrary maps between input and output spaces. The machine scores $\bar{m}$ are the input to a neural network that computes the mapping between them and the corresponding predicted human grade $\tilde{h}$, that is

$$\tilde{h} = o(\bar{m}), \tag{5}$$

where $o(\ )$ represents the nonlinear mapping implemented by the network. For the training of the network, actual human grades in a calibration database provide the targets while the corresponding machine scores provide the input. Using standard training procedures, like *backpropagation*, neural networks will closely approximate Eq. (2), that is, the conditional expected value of the desired output given the inputs [6],

$$o(\bar{m}) \cong E[h|\bar{m}]. \tag{6}$$

Alternatively, if the network has G outputs corresponding to the grade classes, the network outputs will approximate the posterior probabilities $P(h_i|\bar{m})$ needed for classification in Eq. (3) [6].

Trees. Another approach to implementing the mappings is to use classification and regression trees [7]. In our case, a tree can be used to classify a vector of machine scores $\bar{m}$ to one of 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

$$P(h|\bar{m}) \cong P(h|t) \tag{7}$$

where t is the leaf corresponding to the machine scores $\bar{m}$. 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 estimates of the conditional distribution (7) as well as estimates of the expectation (2) can be obtained using this data.

Distribution Estimation. Using this approach we obtain the posterior probabilities needed in Eqs. (3) and (4) by estimating the class conditional distributions of machine scores for each grade $P(\bar{m}|h)$. Then, by using Bayes rule we express $P(h_i|\bar{m})$ as

$$P(h_i|\bar{m}) = \frac{P(\bar{m}|h)P(h_i)}{\sum_{j=1}^{G} P(\bar{m}|h_j)P(h_j)}, \tag{8}$$

where $P(h_i)$ is the estimated prior probability of the human grade h_i. $P(\bar{m}|h_i)$ is modeled by discrete distributions that are estimated based on the quantization of the machine scores. When more than a machine score is used, the use of vector quantization (VQ) allows us to obtain robust estimates of the joint distribution of machine scores, which is modeled by a single discrete distribution on the VQ index $V(\bar{m})$, that is, $P(\bar{m}|h) \cong P(V(\bar{m})|h_i)$.

4. EXPERIMENTS

We give an overview of the speech database used to train and evaluate the scoring models, and then we show results using alternative mappings from machine scores to human grades.

4.1. Training and calibration database

The acoustic models used to generate the phonetic alignments and produce the posterior scores were gender independent, Genonic Gaussian mixture models introduced in [1]. These models were

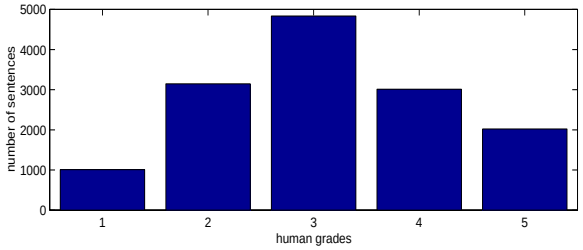

Figure 1: Histogram of human grades for the nonnative spanish sentences.

trained using a gender-balanced database of 142 native Latin American Spanish speakers, totaling about 32,000 sentences.

For the pronunciation scoring experiments we used a database that included 206 nonnative speakers whose native language was American English. The speech material consisted of 14,000 read newspaper sentences. All the speech was recorded in standard offices with computers running, using a high-quality Sennheiser microphone. A panel of five raters, native Spanish speakers, rated the overall pronunciation of each nonnative sentence on a scale of 1 to 5, ranging from "strongly nonnative" to "almost native". The resulting distribution of sentence grades is shown in Figure 1 These human grades were used both to evaluate the effectiveness of the different machine scores as predictors of the pronunciation quality, and to calibrate the mappings from the machine scores to the predicted pronunciation grades. To assess the consistency of these human scores, the correlation between raters was computed in a subset of 2800 sentences that were rated by all five raters. The average sentence/speaker level inter-rater correlation was r=0.68/ 0.91.

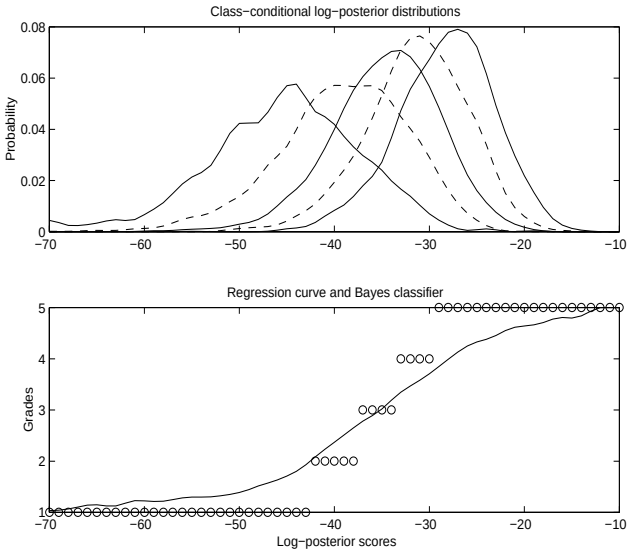

Figure 2: Top: Distribution of log-posterior scores for grade classes 1 to 5 ordered from left to right. Bottom: Mappings for the estimation (continuous trace) and classification (dotted trace) methods for equal priors training

4.2. Evaluation of Mappings

To illustrate some of the characteristics of the mappings we studied the mapping of a single machine score, the log-posterior score, using the density estimation implementation. Similar results can be obtained using the neural networks or tree-based implementations [4]. We scalar quantized the log-posterior scores to estimate the necessary class-conditional probability distributions. For consistency with our design goal of discrete human grades, in the estimation approach the mapped grade was rounded to the nearest integer. We obtained the mappings by using either the database priors or equal priors in Eq. (8). In Figure 2 we show the log-posterior conditional distributions per grade, and the mappings obtained with MMSE estimation (before rounding) and with the Bayes classifier, both for the equal priors case. In the following experiments, the calibration database was divided into two halves, mappings were trained in each half to map the machine scores of the other half, the corresponding results were averaged.

Human-machine correlation	Not mapped	Estimation mapping	Bayes mapping
Evaluated with database priors			
Train map w/database priors	0.59	0.44	0.45
Train map w/equal priors	0.59	0.56	0.55
Evaluated with equal priors			
Train map w/database priors	0.68	0.55	0.56
Train map w/equal priors	0.68	0.66	0.67

Table 1: Human-machine correlations for log-posterior scores unmapped and mapped using MMSE estimation and Bayes classifier.

Effect of the priors. Given that the distributions of machine scores for each grade have significant overlap, the priors have considerable effect in the resulting mappings. For instance, in our case, where the training data for the mappings have the prior distribution shown in Figure 1, the effect of the priors was to widen the region assigned to grade 3 and shift the other boundaries in a way that would make it harder for a speaker to get high or low pronunciation grades. Imposing equal priors over grade classes produced much more consistent grades, with the center of each class region closer to the location of the peaks of the corresponding machine score distributions. This effect can also be assessed quantitatively. In Table 1 we show the human-machine sentence-level correlation for log-posterior scores for unmapped and mapped cases. The correlation was evaluated using the original dataset and also a resampled version with equal priors. We observe that in both evaluations the mappings trained with the original database priors have significantly lower correlations than those trained with equal priors. This lower correlations are due to the fact that the wider regions assigned to the center grade classes quantize more coarsely the machine scores, suppressing part of the correlated variability of the mapped grades. We also notice that even in the equal prior case there is a slight reduction of the correlation compared to the case when no mapping is used. This reduction is attributable again to the quantization of the machine scores in discrete grades, and to the reduction of the dynamic range introduced by the mappings. For

this performance measure, both the estimation and the Bayes classification approaches produced similar results. Also in Table 1, the dependency of the correlation coefficient on the priors of the evaluation database is worth noting

Estimation vs. classification mapping. It is hard to quantify the goodness of either type of mapping, as each one minimizes a different criterion. Therefore, the error measure chosen may favor one or the other. In Table 2 we see that, as expected, the classification error favors the Bayes classifier mapping, while mean square error favors the MMSE estimation mapping. As an alternative to both criteria, we also show in Table 2 the mean absolute error. These results correspond to the case of equal priors for training and evaluation sets. The resulting mean absolute error was similar for both mapping approaches. Nevertheless, from the example in Figure 2 we see that the mappings are different indeed. The estimation mapping is more sensitive to the distribution overlap and has a lower slope than the Bayes mapping. This, in turn, produces a compression of the mapped grades towards the mean range, making harder for a student to obtain the highest or lowest grades. It is easy to show that the slope decreases with the distribution overlap. The Bayes mapping, on the other hand, defines the boundaries between classes based on which is the most likely class, which seems to be a more intuitive criterion for this application; it is also more independent of the distribution overlap.

Error measure	Estimation mapping	Bayes mapping
Classification error	65.5	62.3
Mean square error	1.12	1.19
Mean absolute error	0.80	0.79

Table 2: Different error measures between human and log-posterior scores mapped using estimation and Bayes classifier mappings trained and tested with equal priors.

Extrapolating mappings across languages. After developing scoring systems for two languages, Spanish and French [2], we observed that the relative positions of the grade class boundaries in the mappings were very similar. This occurs in spite of the fact that different human graders have been used in the two systems, and different acoustic models were used to compute the posterior scores. This finding suggests that it could be possible to extrapolate the mapping from one language to another, avoiding the need of a large graded nonnative database for the new language. We propose to use the relative position of the grade class boundaries with respect to the mean or median of the native distribution and a few nonnative scores, to linearly transform the mapping to be applied to a new language. To test this idea, we used an additional database of nonnative French speech produced by american speakers and graded for pronunciation quality [2]. A Bayes classifier mapping was extrapolated from French to Spanish. Using the extrapolated mapping, classification error had a relative increase of 1.2% while human-machine correlation had a relative decrease of 0.3% compared to the case of using the original mapping. These results show very good agreement between the trained and the extrapolated mappings, supporting our previous observation.

5. SUMMARY AND CONCLUSSIONS

We have discussed two calibration methods (MMSE estimation and Bayesian classification) to predict human pronunciation quality grades from machine-generated pronunciation scores. Several nonparametric methods to implement the desired mappings were presented. We evaluated experimentally both approaches and investigated the effects of the database priors in the calibration of the mappings. We concluded that the priors have a strong effect in the mapping of machine scores, and that assuming equal priors produces mappings with higher correlation with the human data. We also argued that the mappings obtained using the Bayes classification approach may have more desirable properties than those obtained using the estimation approach, in terms of better consistency with human data, and in terms of being less affected by the variance of the machine scores. Finally, based on the observation of mappings across different models and languages, we suggested that a simple linear transformation may allow to extrapolate mappings from one language to another.

Acknowledgments

We gratefully acknowledge support from the U.S. Government under the TRP program.

6. REFERENCES

1. 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.

2. 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.

3. 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.

4. H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, “Combination of Machine Scores for Automatic Grading of Pronunciation Quality,” submitted to Speech Communication.

5. H. Franco, L. Neumeyer, and H. Bratt, “Modeling Intra-word Pauses in Pronunciation Scoring,” Proc. of Speech Technology in Language Learning, pp. 87-90, Marholmen, Sweden, 1998.

6. M. Richard, R. Lippman, “Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities,” Neural Computation 3, 461-483, 1991.

7. L. Brahmin, J. Friedman, R. Olsen, and C. Stone, “Classification and Regression Trees,” The Waxworks & Brooks/Cole Statistics/Probability Series, 1984.