AUTOMATIC PRONUNCIATION SCORING OF SPECIFIC PHONE SEGMENTS FOR LANGUAGE INSTRUCTION

Yoon Kim, Horacio Franco, and Leonardo Neumeyer

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

ABSTRACT

The aim of the work described in this paper is to develop methods for automatically assessing the pronunciation quality of specific phone segments uttered by students learning a foreign language. From the phonetic time alignments generated by SRI's Decipher™ HMM-based speech recognition system, we use various probabilistic models to produce pronunciation scores for the phone utterance. We evaluate the performance of the proposed algorithms by measuring how well the machine-produced scores correlate with human judgments on a large database. Of the various algorithms considered, the one based on phone log-posterior-probability produced the highest correlation (r_{xy} = 0.72) with the human ratings, which was comparable with correlations between human raters.

1. INTRODUCTION

An important component of an effective language instruction is the instructor's feedback in the assessment of the pronunciation quality, or detection/correction of specific production problems or common mistakes that a student would make. The recent evolution of speech recognition technology has allowed us to explore new possibilities in computer-aided language instruction, where the computer may provide such feedback.

This work is part of an effort aimed at developing automatic language instruction systems that grade the pronunciation quality of speech uttered by students learning a foreign language [1][2][3]. Previous approaches [1][2] focused on rating an entire sentence rather than targeting specific phone segments. In this paper, we extend previous work by investigating various methods for automatically assessing the pronunciation quality of individual phone segments within a sentence. The ratings obtained may help the student in detecting and/or correcting specific pronunciation problems or mistakes that might serve as obstacles to the improvement of the student's language skills. For a detailed treatment on the problem of automatically detecting mispronunciation, see [4].

In the following, we describe various pronunciation scoring schemes and provide experimental results that evaluate the performance of the algorithms based on how well the machine-produced scores correlate with the corresponding human scores. We also conduct an experiment to estimate the number of phone utterances the student must put into the system to get reliable feedback on his or her pronunciation proficiency.

2. THE DATABASE

The automatic scoring system devised in this paper was developed to help American adults learn the French language [3]. A database of transcribed native read speech was used for training models for speech recognition and pronunciation scoring. A database of nonnative read speech was transcribed and scored for pronunciation quality by expert human raters. The entire corpus consisted of speech recorded from 100 natives of Parisian French (native corpus) and from 100 American students speaking French (nonnative corpus). Speech was recorded in quiet offices by using a high-quality Sennheiser microphone.

A panel of five French teachers rated the pronunciation of selected phone segments (4656 total) in the nonnative corpus. A total of 10 phones (/an/, /eh/, /eo/, /eu/, /ey/, /in/, /on/, /r/, /uw/, /uy/) were considered, and the scores were on a scale of 1 (unintelligible) to 5 (native-like). The raters listened to the sentence containing the phone segment of interest, and were instructed to only consider that particular segment and disregard all other segments in the sentence in giving the scores. They were advised to reject utterances in which the student experienced serious disfluencies or those in which the audio quality was unacceptable. In addition, the sessions were designed such that the raters encountered some of the phone utterances more than once without being informed. This was done to check the self-consistency of a rater.

3. CONSISTENCY OF HUMAN RATINGS

The ratings obtained from the human instructors serve as the ideal target for which the machine-produced ratings should aim. Therefore, it is important to first test the consistency of these human scores both between raters and individually within each rater. For measuring the consistency between scores we used the correlation coefficient.

Using the scores of phone utterances that were rated by all five raters, we computed the correlation between the scores of two raters (*inter-rater correlation*). Two types of inter-rater correlation were computed: at the *phone level*, pairs of corresponding ratings for all the individual phone utterances were correlated; at the *(phone-specific) speaker level*[1], all the phone scores from each speaker were averaged and then the pairs of these speaker-average scores were correlated. The self-consistency of a rater was assessed by correlating the scores of phone utterances that had been rated twice by the same rater (*intra-rater correlation*). The average inter- and intra-rater correlation values are shown in Table 1.

Corr. Type	Level	Number of scores	Correlation
Inter-rater	**Phone**	144	0.55
Inter-rater	**Speaker**	3250 (32.5 per spkr)	0.80
Intra-rater	**Phone**	153	0.86

Table 1: Average inter- and intra-rater correlations across all phones for five human raters.

Comparing the inter-rater correlation values at the phone level (r_{xy} = 0.55) with that at the sentence level (r_{xy} = 0.65) [2] suggests that raters are less consistent with one another when rating phone segments than when rating sentences. This may be explained by considering that in the phone case, the rater has less information to base the score on than in the sentence case. Correlation at the phone-specific speaker-level is high (r_{xy} = 0.80) and comparable to that at the overall speaker-level (r_{xy} = 0.87) [2]. Intra-rater correlation is surprisingly high (r_{xy} = 0.86), suggesting that the human raters are highly consistent with themselves despite being less consistent with one another. The correlation values in Table 1 will serve as the expected upper bound on the performance of the scoring system.

4. PRONUNCIATION SCORING

The pronunciation scoring paradigm [5][6][7] uses hidden Markov models (HMMs) [8], trained using the database of native speakers, to generate phonetic time alignments of the student's speech. From these segmentations, we use the following probabilistic measures [1] to obtain scores for each phone segment.

4.1. HMM-based Log-likelihood Scores

For each phone segment, the log-likelihood score $\hat{l}$ is defined as

$$\hat{l} = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log p(y_t|q_i) \quad (1)$$

where $p(y_t|q_i)$ is the likelihood of the current frame with observation vector y_t, d is the duration (in frames) of the phone segment, and t_0 is the starting frame index of the phone segment. Dividing by d allows us to eliminate the dependency of the pronunciation score on the duration of the phone.

4.2. HMM-based Log-posterior Probability Scores

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:

$$P(q_i|y_t) = \frac{p(y_t|q_i)P(q_i)}{\sum_{j=1}^{M} p(y_t|q_j)P(q_j)} \quad (2)$$

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. The posterior score $\hat{\rho}$ for the phone segment is then defined as

$$\hat{\rho} = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log P(q_i|y_t) \quad (3)$$

4.3. Segment Duration Scores

The procedure to compute the duration-based phone score is as follows. First, from the Viterbi alignment we measure the phone duration in frames; 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 were previously trained from alignments generated for the native training data.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the scoring algorithms we used a test set with an average of 30 sentences from 100 adult American speakers with various levels of proficiency in French.

1.As opposed to *overall* speaker-level correlation, which is obtained by averaging sentence scores [1].

5.1. Human-machine Correlation for Phone Scores

As in the case of assessing the consistency of human raters, two types of correlation were computed between the machine scores and the corresponding human scores (phone level and phone-specific speaker level).

Table 2 shows the phone-level correlations between human and machine scores for some of the individual phone classes considered in this study. An average of about 450 phone scores was used to compute the correlation for each phone class. We see that the measure based on posterior probability is the best (r_{xy} = 0.44) at capturing the pronunciation quality of a single phone segment. Note also that correlation value depends on the phone class, yielding a maximum for /uy/ (r_{xy} = 0.54) and a minimum for /uw/ (r_{xy} = 0.32). Phone duration scores seem to be almost uncorrelated with the corresponding human ratings (r_{xy} = 0.06), and thus duration turns out to be a very poor measure at the phone level. This could be understood by observing the following: First, phone duration for vowels is highly variable even among native speakers. Second, only a single measurement of phone duration is used to score a phone. So unlike the other two features, there is no segment averaging of measured features, making duration a noisy feature at the phone level.

Phone	Duration	Likelihood	Posterior
/an/	0.13	0.23	0.40
/in/	0.08	0.29	0.45
/r/	0.06	0.23	0.48
/uw/	0.07	0.21	0.32
/uy/	-0.03	0.41	0.54
Avg.	0.06	0.27	0.44
Sent.-level Corr.	0.47	0.33	0.58

Table 2: Human-machine correlation at the phone level for various scoring measures. Sentence-level human-machine correlation values from [1] are also shown at the bottom for comparison.

Table 3 shows correlations between human and machine scores at the speaker level with a total of 4656 phone segments across 100 speakers. We also show the overall speaker-level correlation values from [1] for comparison. We see that the correlation values corresponding to specific scores are lower than those in the case of overall scores. This is especially true for duration scores, which reconfirms that phone duration should be used for pronunciation scoring only when there is a sufficient amount of data for averaging. At the speaker level, the posterior-probability-based score again has the highest correlation value (r_{xy} = 0.72).

Algorithm	Level of Correlation	
	Specific Spkr.	Overall Spkr.
Normalized Duration	0.46	0.84
Likelihood	0.36	0.50
Posterior Probability	0.72	0.88

Table 3: Phone-specific speaker-level correlation values between human and machine scores along with overall speaker-level correlation values from [1].

We now compare the human-machine correlation values from the above tables with correlations between human raters that were obtained in Section 3. Figure 1 gives a comparison between the two. As we can see, at the speaker level, the performance of the scoring system (in terms of correlation with human raters) is comparable to that of a human rater. However, at the phone level, the correlation only reached 80% (r_{xy} = 0.44) of the target value (r_{xy} = 0.55) corresponding to correlation between human raters.

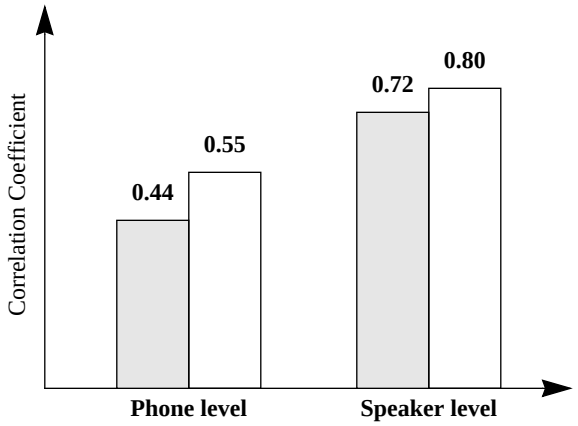

Figure 1: Comparison of human-machine correlation with human-human correlation at the phone and speaker level. Shaded bars represent the average human-machine correlation while white bars represent the average human-human correlation.

5.2. Effect of Varying the Amount of Speaker Data

To evaluate the system's performance (in terms of human-machine speaker-level correlation) as a function of the number of test utterances per speaker, we conducted another experiment. We varied the number of machine phone scores per speaker (N) from 10 to 320 in obtaining the average machine score for each speaker. Then, for each N, we computed the correlation between these speaker-average machine scores and the speaker-average human scores obtained using the *entire* human score data. Of course, the assumption here is that the speaker-average scores from the entire database (4656 scores, 46.56 scores per speaker) of

human ratings approximate the true pronunciation proficiency of the students.

In extracting the machine scores across all 10 phones, constant proportion is maintained for each phone. For example, if the phone /an/ constitutes 20% of the 10 phones uttered in the entire database, and /eh/ constitutes 10%, then for N=40, we would extract $40 \times 0.20 = 8$ scores from /an/ and $40 \times 0.10 = 4$ scores from /eh/, and so on.

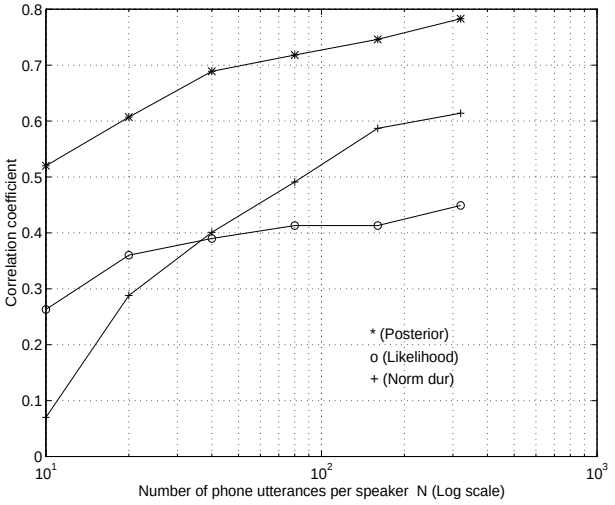

Figure 2: Speaker-level human-machine correlation for various numbers of phone utterances per speaker.

As we can see in Figure 2, we get improvement in performance as we increase the number of scores per speaker N. This is not surprising since more information leads to a more accurate evaluation of the speaker's overall proficiency in pronouncing the phones. Correlation for duration scores increases drastically with increase of data from N=10 to N=160 because of the effect of averaging as discussed earlier. Posterior probability measure outperforms the other measures for all values of N, and its correlation value rapidly increases up to about 0.7 at N=40. This suggests that about 40 phone utterances per speaker is sufficient for obtaining a reliable score from the system.

6. CONCLUSIONS AND FUTURE WORK

We have presented methods to capture the pronunciation quality of specific phone segments. We evaluated the performance of the proposed algorithms by measuring how well the machine scores correlate with corresponding human scores in the case of rating single phone utterances (phone level) and also when judging a student's overall ability to pronounce particular phones over multiple utterances (speaker level). We found that the algorithm based on phone posterior probability produced ratings that have the highest correlation with the human ratings in both cases. At the speaker level, the system's performance was comparable to that of a human rater; however, at the phone level, there exists a performance gap, which calls for further research in the case of rating a single phone utterance.

Research presented in this paper was based on a pilot study with a limited amount of test and training data. We are currently working on the collection of a larger database of human ratings. The new database will focus on phones that have more linguistic importance (e.g., phones that are commonly problematic to students) in addition to having more phone ratings per speaker. This will help us devise new schemes to improve ratings of individual phone utterances. In addition, the scoring algorithms could be combined with algorithms that detect mispronunciation [4] to provide a more comprehensive feedback to the student.

7. ACKNOWLEDGMENT

The authors wish to gratefully acknowledge support from the U.S. government under the TRP Program.

8. REFERENCES

1. H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic Pronunciation Scoring for Language Instruction", *Proc. of ICASSP 97*, pp. 1471-1474, Munich, Germany, 1997.

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. M. Rypa, "ECHOS: A Voice Interactive Language Training System", *Proc. of CALICO*, Alburquerque, New Mexico, 1996.

4. O. Ronen, L. Neumeyer, and H. Franco, "Automatic Detection of Mispronunciation for Language Instruction", *Proc. of Eurospeech 97*, Rhodes, Greece, 1997.

5. J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic Evaluation and Training in English Pronunciation", *Proc. of ICSLP 1990*, Kobe, Japan, 1990.

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

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

8. V. Digalakis and H. Murveit, "GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer", *Proc. of ICASSP 94*, pp. 1537-1540, 1994.