AUTOMATIC TEXT-INDEPENDENT PRONUNCIATION SCORING OF FOREIGN LANGUAGE STUDENT SPEECH

Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price

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

ABSTRACT

SRI International is currently involved in the development of a new generation of software systems for automatic scoring of pronunciation as part of the *Voice Interactive Language Training System* (VILTS) project. This paper describes the goals of the VILTS system, the speech corpus, and the algorithm development. The automatic grading system uses SRI's Decipher™ continuous speech recognition system [1] to generate phonetic segmentations that are used to produce pronunciation scores at the end of each lesson. The scores produced by the system are similar to those of expert human listeners. Unlike previous approaches in which models were built for specific sentences or phrases, we present a new family of algorithms designed to perform well even when knowledge of the exact text to be used is not available.

1. INTRODUCTION

Computer-aided language instruction has been evolving from simple systems with exercises based on text and static pictures to more advanced systems that accept user input text or pointing, and may also involve speech output. More recently, the possibility of accepting speech input began to become practical. The addition of speech input allows developers to complement reading and listening comprehension (receptive skills) with more active 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 key in 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 match the perceptual capabilities of human listeners to grade speech quality.

Previous work at SRI [2, 3, 4] used speech recognition technology to score the pronunciation of Japanese students speaking English over the telephone based on fixed text prompts. Knowledge of the text can be used to compute robust pronunciation scoring algorithms, but limits generalizability, since new lessons will require additional data collection. We refer to this class of algorithms as *text-dependent* because they rely on statistics related to specific words, phrases, or sentences. Measures related to the likelihood of segmental spectral features and duration were found to correlate very well with human ratings.

Recently SRI started development of the VILTS project [5] to incorporate spoken language technology in a system geared toward training foreign language students. The first version of the system was designed to teach French to students whose first language is American English. The system elicits speech through various language instruction activities designed to ensure that the recognizer produces a correct transcription of the recordings 99% of the time. This transcription is used to produce an accurate phonetic segmentation used by the system to produce pronunciation scores that correlate well with those of expert human listeners.

The VILTS software is designed to be extensible and flexible; language instructors should be able to modify and design lessons without expert knowledge in speech recognition technology. To achieve this goal, we developed text-independent pronunciation scoring algorithms. To develop the algorithms, an extensive speech corpus was designed and collected.

2. THE VILTS CORPUS

The VILTS project required data for speech recognition, for pronunciation algorithm development, and to provide core lesson material. Speech was recorded from 100 natives of French living in Paris, strong regional accents were avoided. We refer to this data as the *native* corpus. The *nonnative* corpus was recorded from 100 American students speaking French. The speech was recorded in quiet offices using a high-quality Sennheiser microphone. The natives were recorded in four modes:

- Read speech, common sentences, designed to include most common pronunciation problems for American students;
- Read speech, newspaper sentences, which were not read within the native speaker corpus by more than one speaker;
- Spontaneous conversations between a subject and an interviewer; and
- Read speech versions of the conversation transcripts by the same speakers.

The nonnative corpus consisted of:

- Read speech, common sentences (same sentences used in the native corpus);
- Read speech, newspaper sentences; and

- Read/imitated speech, in which the subject was able to listen to a native reading the same sentence before starting the recording.

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. About 10% of the data was rated by all five teachers and twice by each teacher. Multiple ratings of the same utterance were used to evaluate inter- and intra-correlations among the raters.

Pronunciations for French words used in the corpus were generated by a text-to-speech system and revised by a linguist. 37 phonemes were used, and each word could have multiple pronunciations (French liaison was modeled using multiple pronunciations).

3. PRONUNCIATION SCORING

Human scores are the reference against which the performance of the scoring systems is validated. For this reason it is important to asses the consistency of human scores, both between raters and within each rater. To measure human consistency and to evaluate automatic scores we use simple linear correlation techniques.

3.1. Human Scoring

Human judgments were provided by the five raters of speech from the 100 students. Using the subset of sentences scored by all raters, we assessed inter-rater correlation based on individual sentence scores and on individual speakers (Table 1).

Rater	1	2	3	4	5
1	1.00/1.00	0.61/0.84	0.68/0.75	0.67/0.79	0.70/0.85
2		1.00/1.00	0.60/0.79	0.55/0.74	0.60/0.82
3			1.00/1.00	0.66/0.75	0.70/0.82
4				1.00/1.00	0.72/0.86
5					1.00/1.00

Table 1: Sentence/Speaker-level correlations between raters

The level of correlation is reasonably uniform among the pairs of raters. The correlations at the speaker level are consistently higher than those at the sentence level, reflecting that the average scores based on several sentences are more reliable than the scores based on single sentences. The average correlation between raters at the sentence level is 0.65 while at the speaker level it reaches 0.8. We also computed the correlation between a rater and the mean of all other raters excluding the current one. Table 2 shows this type of correlation at the sentence level and speaker level. This way of assessing the correlation among raters at the speaker level is similar to the way the machine scores will be correlated with human scores. Correlation between a rater and a pool of other raters also suggests an upper bound on the level of correlation between human and machine scores. Table 2 also shows the intra-rater correlation, assessing the consistency of repeated judgments of the same material by the same rater. In particular, each rater was asked to rate the same utterance twice, on different days and in different contexts. As we would expect, comparing with Table 1, the intra-rater correlation is higher than the average of pair-wise inter-rater correlation (0.65), reaching an average of 0.76.

Correlation Type	Level	Rater Ids					Avg.
		1	2	3	4	5	
Inter-rater	Sent	0.78	0.67	0.77	0.76	0.80	0.76
Inter-rater	Spkr	0.88	0.86	0.84	0.85	0.92	0.87
Intra-rater	Sent	0.82	0.73	0.86	0.71	0.75	0.76

Table 2: Sentence- and speaker-level correlations. Inter-rater correlations are computed against the average of the other raters. Intra-rater correlations are computed using two ratings of the same utterance by the same rater.

Descriptive statistics were obtained over the whole set of almost 20,000 human scores of nonnative data from 100 speakers. The histogram of the scores, using a scale from 1 to 5 described earlier, from all raters for all sentence types is shown in Table 3.

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

Table 3: Histogram of scores across all sentence types and raters.

We note a smaller number of level-5 ratings, consistent with the fact that these are ratings for nonnatives. The maximum of the distribution is for the score 3, and shows a significant asymmetry toward lower scores. In Table 4, the mean and standard deviation of the scores given by each rater are shown. The means differ at most by a half point, and the standard deviations are reasonably similar.

Rater ID	1	2	3	4	5	Avg.
Mean	2.5	2.7	3.0	2.5	3.0	2.7
Std. Dev.	0.8	0.8	0.9	0.9	1.1	0.9

Table 4: Means and standard deviations of scores from each rater.

Table 5 shows the average scores for each sentence type. The average score correlates well with the level of difficulty of the task (read sentences are more difficult than imitated sentences, and newspaper sentences more difficult than common sentences).

Sentence Type	Mean
Common Sentences Imitated	3.0
Newspaper Sentences Imitated	2.7
Common Sentences Read	2.8
Newspaper Sentences Read	2.5

Table 5: Means of scores for each sentence type.

3.2. Automatic Scoring

We developed various pronunciation scoring algorithms that rely on phonetic time alignments produced by SRI's speech recognition system. To generate the alignments, we must recover the text read by the student. We do this by eliciting speech in a constrained way in the language learning activities. The algorithms were designed according to the following objectives: (1) machine scores must correlate well with human expert listener scores and (2) no statistics of specific phrases or sentences should be used (i.e., the algorithms must be text-independent). Algorithms in four categories were investigated: hidden Markov model (HMM) log-likelihood scores, segment classification scores, segment duration scores, and timing scores. Each of these categories of scores is described below.

3.2.1. HMM Log-Likelihood Scores

In this approach, we use the HMM log-likelihood as scores. 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 between native speech and nonnative speech. For each sentence, the phone segmentation is obtained, along with the corresponding log-likelihood of each segment. However, for a given level of mismatch between speech and models, with the standard assumptions in the HMM framework, the log-likelihood depends on the length of the sentence. To normalize for the effect of the sentence length we use the "global average log-likelihood" score [4], defined as:

$$G = \left(\sum_{i=1}^{N} l_i\right) \Big/ \left(\sum_{i=1}^{N} d_i\right)$$

where l_i is the log-likelihood corresponding to the ith phone and d_i is its duration in frames, with sums over the number of phones. The degree of match during longer phones tends to dominate the global log-likelihood score. Although shorter phones may have an important perceptual effect, as their duration is smaller, the degree of mismatch along them may be swamped by that of longer phones. To attempt to compensate for this effect we use the following "local average log-likelihood" score L [4], defined as:

$$L = \frac{1}{N}\sum_{i=1}^{N} \frac{l_i}{d_i}$$

where the variables are defined as above. In this score, the degree of match for each phone is weighted equally regardless of its length.

3.2.2. Segment Classification Scores

Another approach to assessing pronunciation is to compute phone classification error; if the phone classifier is trained using native speakers, then the closer the test speaker is to the training population, the higher the classification accuracy should be. We implemented a French phone recognizer and used recognition accuracy as a pronunciation score.

3.2.3. Segment Duration Scores

Relative phone duration should correlate well with the human expert listener's scores for psychological and linguistic reasons. The cognitive load of thinking about how to articulate can disrupt the speech flow and increase disfluency. Cross-language differences a nonnative may impose from the native language on the language being learned can also affect durations of segments. Differences in letter to sound rules for the orthographies of the two languages may lead to insertions, deletions or substitutions of phones that will result in duration differences. Since, to achieve text independence, we cannot use sentence, phrase, or word durations to normalize phone durations, we use a measure of rate of speech (ROS) as the normalization factor. The simplest approach to ROS is to compute the global rate of speech as the average number of phones per unit of time for a given speaker. Normalized duration can be computed as $\bar{d}_i = d_i \cdot ros_s$ where d_i is the unnormalized duration for segment i and ros_s is the estimated rate of speech for speaker s. To compensate for phone alignment errors near silence, we investigate the effect of excluding phones in the context of silence from the train and test data sets.

3.2.4. Timing Scores

Insofar as nonnative speakers tend to speak more slowly than natives, speaking rate should be a good predictor of fluency and can be used as a pronunciation score. Other aspects of linguistic timing can also be exploited since language learners tend to impose the rhythm of their native language on the language they are learning. For example, English tends to be *stress-timed* (stressed syllables tend to be lengthened and others shortened), while Spanish and French tend to be *syllable-timed*. In our investigations a distribution of normalized syllabic periods is computed between the centers of vowels within segments of speech. The normalized time between syllables is used to produce a syllabic timing score.

3.3. Experimental Results

To evaluate the pronunciation scoring algorithms, we used a test set with an average of 30 common sentences from 100 adult American speakers with various levels of proficiency in French. The recordings were verified by the human expert listeners at the same time that they rated the pronunciations. Listeners were instructed to reject utterances in which the audio was contaminated during the recording and those in which the student was seriously disfluent, stumbled, or had other significant disruptions. A French recognizer was trained using SRI's Decipher™ speech recognition system [1]. We used 16,000 utterances from 100 native speakers reading newspaper text. Phone recognition performance was evaluated using 37 phonetic classes with a bigram phone model; phone recognition error rate on this task was 20.6%. We report (Table 6). correlations between machine and human scores computed at the sentence level (across 3000 sentences) and speaker level (across 100 speakers).

To compute native statistics for the pronunciation algorithms and to evaluate the correlation between human and machine scores, we generated phonetic time alignments for all the native and nonnative data using the Viterbi decoder.

Both global and local HMM likelihoods are very poor predictors of pronunciation ratings. It is not clear why in the global likelihood score, correlation decreases when the silence is excluded (A1 vs. A2). The opposite effect can be observed for the local likelihood scores (A3 vs. A4). Phone classification results in similar performance at the speaker level but seems to correlate better at the sentence level. Segment duration scores produce the best results at the speaker level. Normalizing duration helps (C1 vs. C2) and should also increase robustness, as the scores become independent of the rate of speech. Nonparametric distributions also improve performance compared to the single Gaussian case (C2 vs. C3).

Exp	Algorithm	Correlation	
		Sent	Spkr
	A. HMM Scores		
A1	Global log-likelihood, with silence	0.276	0.429
A2	Global log-likelihood, no silence	0.182	0.313
A3	Local log-likelihood, with silence	0.255	0.406
A4	Local log-likelihood, no silence	0.285	0.481
	B. Segment Classification Scores		
B1	Phone recognition	0.399	0.469
	C. Segment Duration Scores		
C1	Duration (Single Gaussian per phone)	0.463	0.735
C2	Normalized duration (Single Gaussian per phone)	0.452	0.827
C3	Normalized duration (Discrete distributions)	0.453	0.845
C4	Normalized duration (Discrete distributions, no silence context)	0.410	0.856
	D. Timing Scores		
D1	Global rate of speech	0.408	0.685
D2	Normalized syllabic timing	0.355	0.726

Table 6: Sentence and speaker level correlations between human and machine scores using 100 nonnative speakers and 30 utterances per speaker.

This improvement is not surprising since the probability distribution of phone duration is not Gaussian. Excluding phones in the context of silence produces a small improvement in correlation at the speaker level (C4 vs. C3). Sentence-level results are still poor, suggesting that further work is needed to predict pronunciation ratings using only a single utterance.

Finally, the timing scores result in acceptable speaker level correlations. Global rate of speech is a good predictor of pronunciation rating, confirming that advanced students speak faster than beginners. However, this score by itself would be a poor indicator of overall pronunciation given that any speech-like signal of the right duration could result in high machine scores. Syllabic timing, however, should be robust to ROS because the durations are normalized and affected only by the relative duration of the timing between syllables.

To evaluate the correlation as a function of the amount of test data, we conducted a second experiment. In this case, we used various amounts of newspaper text from all 100 nonnative speakers to compute the correlations. The results are shown in Table 7.

Number of Sentences	A4	C4	D2
01	0.382	0.512	0.420
05	0.488	0.759	0.669
10	0.490	0.779	0.657
20	0.509	0.816	0.711
30	0.502	0.815	0.712
40	0.493	0.817	0.714
50	0.503	0.830	0.720

Table 7: Speaker-level correlation for various amounts of test sentences using three different methods

Clearly, correlations improve as the amount of test data increases. At least five sentences appear to be required to produce reasonable pronunciation scores. Spectral scores (A4) seem to be more erratic than duration (C5) and timing scores (D2). Duration scores produce the best correlation in all cases.

4. SUMMARY

We have presented the algorithms being developed to generate reliable pronunciation scores. We compared different methods and found that those based on normalized duration scores produced the best results. This finding indicates that relative phone duration is a good predictor of pronunciation proficiency. Moreover, duration scores should be more robust to stressed conditions such as background noise or limited channel bandwidth than are pure spectral scores.

5. REFERENCES

1. V. Digalakis and H. Murveit, "GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer", *IEEE ICASSP, 1994, pp. 1537-1540.*
2. J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic Evaluation and Training in English Pronunciation", *ICSLP 1990, Kobe, Japan.*
3. J. Bernstein, "Automatic Grading of English Spoken by Japanese Students", *SRI International Internal Reports Project 2417, 1992*
4. V. Digalakis, "Algorithm Development in the Autograder Project", *SRI International Internal Communication, 1992.*
5. M. Rypa, "VILTS: The Voice Interactive Language Training System", *to appear in Proceedings of CALICO 1996.*