ELSEVIER

Speech Communication 30 (2000) 83–93

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Automatic scoring of pronunciation quality

Leonardo Neumeyer *, Horacio Franco, Vassilios Digalakis, Mitchel Weintraub

SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA

Received 30 January 1998; received in revised form 25 November 1998; accepted 15 January 1999

Abstract

We present a paradigm for the automatic assessment of pronunciation quality by machine. In this scoring paradigm, both native and nonnative speech data is collected and a database of human-expert ratings is created to enable the development of a variety of machine scores. We first discuss issues related to the design of speech databases and the reliability of human ratings. We then address pronunciation evaluation as a prediction problem, trying to predict the grade a human expert would assign to a particular skill. Using the speech and the expert-ratings databases, we build statistical models and introduce different machine scores that can be used as predictor variables. We validate these machine scores on the Voice Interactive Language Training System (VILTS) corpus, evaluating the pronunciation of American speakers speaking French and we show that certain machine scores, like the log-posterior and the normalized duration, achieve a correlation with the targeted human grades that is comparable to the human-to-human correlation when a sufficient amount of speech data is available.

Keywords: Automatic pronunciation scoring; Speech technology; Hidden Markov models; Speech recognition; Pronunciation quality assessment; Language instruction systems; Computer aided language learning

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 the more active processes of production and conversation. In these systems, the computer provides 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 to the automatic evaluation of pronunciation quality. 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.

Initial work at SRI International (Bernstein et al., 1990; Bernstein, 1992; Digalakis, 1992) 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 this approach limits generalizability, since new lessons

* Corresponding author.

0167-6393/00/$ - see front matter
PII: S0167-6393(99)00046-1

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.

More recently, in the Voice Interactive Language Training System (VILTS) project (Rypa, 1996; Neumeyer et al., 1996; Franco et al., 1997), researchers at SRI incorporated 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 elicited speech through various language instruction activities designed to ensure that the recognizer produced a correct transcription of the recordings 99% of the time. This transcription, and its corresponding phonetic segmentation, were used by the system to produce pronunciation scores that correlated well with those of expert human listeners. To ensure that the VILTS software was extensible and flexible and that language instructors would be able to modify and design lessons without expert knowledge in speech recognition technology, the original text-dependent scoring algorithms were extended and generalized and new algorithms were devised so that text-independent pronunciation scoring was possible (Neumeyer et al., 1996; Franco et al., 1997).

A significant part in the overall design work of a system that evaluates the pronunciation quality of foreign-language students should be devoted to the interface and the selection of text material, so that the overall interaction is adapted to the student level (Bernstein, 1992; Rypa, 1996; Neumeyer et al., 1996). In this paper, however, we focus on the main algorithmic issues of the automatic evaluation of pronunciation. In Section 2 we present the standard pronunciation scoring paradigm that has been used in (Bernstein et al., 1990; Digalakis, 1992; Neumeyer et al., 1996; Franco et al., 1997); in this paradigm, the grades of expert human raters are used to calibrate the machine scores. The collection and validation of human scores is discussed in Section 3. The machine scores used for the automatic evaluation of pronunciation are presented in Section 4 and are evaluated in Section 5 on the VILTS task. Our conclusions are given in Section 6.

2. Pronunciation scoring paradigm

In previous work on the automatic evaluation of pronunciation (Bernstein et al., 1990; Digalakis, 1992; Neumeyer et al., 1996; Franco et al., 1997), a common framework has been adopted in which read or spontaneous speech is first elicited from the student, according to the student's fluency in the foreign language. In the second stage, the student's pronunciation is compared to the pronunciation of native speakers, using native-speaker data collected for this purpose. A machine score is assigned in the third and final stage, to predict the grade that expert human listeners would assign to the student for the specific skill that is being examined.

The pronunciation scoring paradigm developed in (Bernstein et al., 1990; Digalakis, 1992; Neumeyer et al., 1996; Franco et al., 1997) uses hidden Markov models (HMMs) (Digalakis et al., 1996) to recover the text read by the student and to generate *phonetic segmentations* of the student's speech, identifying the start and end times of the different phones spoken in a sentence by the student. With these segmentations, spectral match and prosodic scores can be derived by comparing the student's speech to the speech of native speakers. The generation and calibration of machine scores follows three main steps:

1. Generation of a *phonetic segmentation*, using an HMM-based speech recognizer. The recognizer models can be trained on data from both native speakers and nonnative students of the foreign language.
2. Creation of *machine pronunciation scores* for the different phonetic segments by comparing the speech of the student to that of native speakers.
3. *Calibration* of the scores, which includes tuning the machine scores and possibly combining several of them. The goal is to develop scores that match as closely as possible the judgment of expert human listeners. To achieve this, it is necessary to collect training data that include pronunciation ratings by expert human raters.

The first two steps can be integrated, as proposed in (Ronen et al., 1997). There, the recognition step is performed using a network that models mispronunciation, and machine scores are assigned based on the path that is followed in the network. In this paper, however, we focus on machine scores developed using the procedure that we outlined above.

Several resources are required to evaluate the pronunciation with the procedure described here. These resources typically include a corpus of native speech data, a corpus of nonnative speech data and a corpus of human ratings for the pronunciation skills that the system will be asked to evaluate.

2.1. Native speech corpus

The native speech data are used to train the speech recognizer that generates the phonetic alignments of the student's speech, and to provide a reference for pronunciation quality. Therefore, the population characteristics of native speech data must match as closely as possible the characteristics of the nonnative-student target population. This minimizes noise in the phonetic alignments and in the pronunciation scoring due to speaker mismatch. In (Bernstein, 1992), where the target population consisted of Japanese students speaking over the telephone, the native corpus consisted of American and British students whose age was estimated to correspond to the age of the targeted Japanese students.

The text material of the native corpus depends on the mode of operation of the pronunciation-evaluation system (text-dependent or independent). In (Bernstein, 1992), where lessons with fixed-text prompts were used, the native students recorded the text on which the Japanese students were going to be evaluated. In the VILTS project (Neumeyer et al., 1996), however, where the flexibility of text-independent pronunciation evaluation was desired, the text and the recording procedure of the native corpus were designed so that the native corpus would be useful for different skill levels of the nonnative students and would facilitate the creation of new lessons with new text. Hence, the native speakers were recorded in four modes:

- Read speech, common sentences, designed to include most common pronunciation problems for nonnative students.
- Read speech, newspaper sentences, which were not read within the native speaker corpus by more than one speaker, so that many different words would be collected.
- Spontaneous conversations between a subject and an interviewer, that could be used at higher skill levels.
- Read speech versions of the conversation transcripts by the same speakers.

2.2. Nonnative speech corpus

The nonnative speech corpus can be used in many ways. Part of the nonnative speech data may be combined with the native speech data and used to train the recognizer that will align the student's speech. This approach makes the recognizer more robust to common pronunciation mistakes made by nonnative students. A second part of the nonnative corpus may be used for development and calibration of machine pronunciation scores, whereas a third part is typically used for evaluation and validation of scoring algorithms.

The parts of the nonnative corpus that are used for development and evaluation of machine scoring are complemented by grades from human experts (teachers or native speakers) for the various skills that are examined. The machine scoring problem can then be defined as the prediction of the grade that an expert human listener would assign to the pronunciation skills. The development part of the corpus with its corresponding human grades can be used as training data to estimate the model parameters, that is, the predictors used by the machine to estimate the human grade. The validity of these predictors is tested on the evaluation part of the nonnative corpus.

For text-independent scoring, the nonnative corpus design guidelines are similar to those of the native corpus. As an example, the nonnative corpus of the VILTS project consisted of

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

- Read and imitated speech, in which the subject listened to a native speaker reading the same sentence before starting the recording.

2.3. Human ratings corpus

The human ratings corpus is created by providing multiple human judgments for various pronunciation skills and for all the sentences in the development and validation parts of the nonnative speech corpus. The performance measure that has been typically used for both the development and validation phases is the correlation between the machine scores and the corresponding human ratings (Bernstein et al., 1990; Digalakis, 1992; Neumeyer et al., 1996; Franco et al., 1997) and in this paper we adopt the same performance measure. It is, however, important that the human ratings, the reference against which the performance of the scoring system is validated, are consistent both within and across raters.

3. Human scoring

Human assessment of pronunciation is a highly subjective task. The assessment of human grading in the VILTS project was based on previous work by Bernstein (1992) and Greenberg et al. (1992) for the Autograder project. In Greenberg's study, Japanese students were asked to read sentences over the telephone. Each utterance was later graded on a 1–7 scale. A rating of 7 indicated native-like pronunciation, and a rating of 1 indicated that the utterance had a strong foreign accent and was difficult to understand. Twenty-nine speakers of American English evaluated almost 60,000 utterances. Although more than one listener rated the same utterance, there was no common set of utterances rated by all the raters. Inter-rater reliability was computed on a subset of the sentences rated by at least 10 listeners. The average pair-wise inter-rater correlation was on the order of $r = 0.55$.

In the VILTS study we made a few changes to the experimental design in consultation with Steve Greenberg. We decided to use a 1–5 scale instead of 1–7, because the latter was more likely to introduce inconsistencies in the ratings than to add granularity to the results. To control the consistency of the ratings, the judges were asked to rate a subset of the materials (about 10% of the nonnative data) more than once. This control subset was used to estimate the intra- and inter-rater reliability systematically throughout the study. We began by conducting a pilot study with 10 expert human listeners. Based on the results of the pilot, we selected the five most consistent raters, who were used to judge utterances from all 100 students. The measures used to evaluate consistency within and across speakers were intra- and inter-rater correlation, respectively, as explained below.

In Table 1 we summarize the inter-rater correlation on the control subset. We computed the correlations at the sentence- and speaker-levels. Sentence-level inter-rater correlation between two raters A and B is computed by evaluating the correlation coefficient between the grade given by rater A and that of rater B for all utterances in the common set. To compute the speaker-level inter-rater correlation between two raters A and B, the average score of each rater for all the utterances of each speaker is first calculated. These speaker-level scores of raters A and B are then used to calculate the correlation coefficient.

Table 1
Sentence/speaker-level correlations between raters

Rater Id	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 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

		Rater Ids					
Correlation type	Level	1	2	3	4	5	Average
Inter-rater	Sentence	0.78	0.67	0.77	0.76	0.80	0.76
Inter-rater	Speaker	0.88	0.86	0.84	0.85	0.92	0.87
Intra-rater	Sentence	0.82	0.73	0.86	0.71	0.75	0.76

Table 3
Histogram of scores across all sentence types and raters

Score	1	2	3	4	5
Percentage (%)	9	31	42	15	3

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. Based on previous studies, we believe that this level of correlation is acceptable within the limitations of the experimental design. We also computed the correlation between a rater and the mean of all other raters excluding the current one, also referred to as *open correlation*. The open correlation at the sentence level is computed by evaluating the correlation coefficient between the grade of a particular rater and the mean of the grades of the remaining raters over all sentences in the common set. The open correlation at the speaker level is computed by first averaging each of these two scores over all sentences coming from the same speaker. Table 2 shows results for the open 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. Hence, 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.

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

Table 5 shows the average scores for each sentence type in the VILTS corpus. The average score correlates well with the level of difficulty of the task (read sentences are more difficult than

Table 4
Means and standard deviations of scores from each rater

Rater Id	1	2	3	4	5	Average
Mean	2.5	2.7	3.0	2.5	3.0	2.7
Standard deviation	0.8	0.8	0.9	0.9	1.1	0.9

Table 5
Means of scores for each sentence type

Sentence type	Mean
Common sentences imitated	3.0
Newspaper sentences imitated	2.7
Common sentences read	2.8
Newspaper sentences read	2.5

imitated sentences and newspaper sentences more difficult than common sentences).

4. Machine scoring

The machine scores must be reliable and must correlate well with the human expert listener scores. To achieve this, several issues must be considered.

- The system must be protected from recognition errors. If the degree of spectral and prosodic match to the speech of native speakers is used to create machine scores and there are errors in the phonetic segmentations, then the student's voice will be compared to the incorrect templates, thereby introducing significant noise in the pronunciation scores. Hence, speech in automatic language learning activities must be elicited in a constrained way. The constraints should depend on the student's level, since the recognizer's performance is directly related to the speaker's fluency. Beginners may be prompted to read from specified text, as in (Bernstein, 1992; Digalakis, 1992); as the student's level advances, multiple-choice questions may be used (Rypa, 1996).
- Good machine scores must measure only the student's ability in the pronunciation of the language's phones and prosody and not the degree of closeness between the student's voice and the native-speaker voices used as a reference. The machine scores must, therefore, be normalized against such factors as the degree of acoustic match between the student and the reference native speakers (useful in spectral match scores), and the rate with which the student speaks (useful in prosodic scores).
- The amount of speech, upon which the evaluation of the pronunciation is based, must be long enough to ensure reliable scores.

In the remainder of this discussion, we assume that the interaction between the student and the machine has been designed to allow error-free recognition of what the student has said. If we assume that we know what the student is saying, as in the case of read speech, we can obtain forced Viterbi alignments using the model of the known sentence. Hence, by using a speech recognizer we can obtain a fairly reliable phonetic segmentation of the student's speech, which includes the labels and start and end times of the phones present in the student's response to the prompting material. To address reliability issues related to the length of speech, we present scores that are obtained over either a single utterance or a group of utterances (a *lesson*) and provide a measure of the student's overall pronunciation quality. An overall assessment of the pronunciation quality is something that can be estimated reliably in that time frame. Most of the scores can also be used to evaluate the pronunciation of specific phones, albeit at a longer time frame (Kim et al., 1997). The issue of speaker normalization is, however, an inherent part of the individual scores and will be addressed separately in Sections 4.1–4.5, which describe five categories of algorithms: HMM log-likelihood scores, normalized acoustic scores, segment classification scores, segment duration scores, and timing scores.

Machine scores can be classified as text-dependent or text-independent. Text-dependent scoring algorithms use information about the text to produce and calibrate the machine scores. When the lessons that are used to evaluate the pronunciation are fixed, as in (Bernstein, 1992; Digalakis, 1992), then statistics of the specific words and phrases can be used to normalize the machine scores, combine the sentence-level scores into a single measure of pronunciation quality, and so forth. However, text-dependent scoring algorithms make the development of new lessons cumbersome, because additional data from native speakers and nonnative students must be collected, since they are needed to create and calibrate the machine scores. Hence, we have recently focused on the harder problem of creating text-independent pronunciation scores (Neumeyer et al., 1996; Franco et al., 1997).

4.1. HMM log-likelihood scores

In this approach, we compute, using the HMM, the log-likelihoods of spectral observations extracted from short-time windows of speech (frames) and use these log-likelihoods as scores. The underlying assumption is that the logarithm of the likelihood of the speech data, computed by the Viterbi algorithm, is a good measure of the similarity between native speech and nonnative speech when the HMMs are trained using native speech data. For each sentence, the phone segmentation is obtained, along with the corresponding log-likelihood of each segment. Let τ_i denote the start time of the ith phonetic segment. The total log-likelihood of this segment can be computed, using an HMM, by

$$l_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \log(\mathrm{p}(s_t|s_{t-1})\mathrm{p}(x_t|s_t)), \tag{1}$$

where x_t, s_t denote the observed spectral vector and the HMM state at time t, respectively, $\mathrm{p}(s_t|s_{t-1})$ is the HMM transition probability and $\mathrm{p}(x_t|s_t)$ is the so-called output distribution of state s_t. Since this log-likelihood depends on the length of the sentence, we must normalize for the effect of the sentence length. In (Digalakis, 1992) two methods were proposed for this. The first, the "global average log-likelihood" score G, is defined as

$$G = \frac{\sum_{i=1}^{N} l_i}{\sum_{i=1}^{N} d_i}, \tag{2}$$

where the summations are over all N segments in an utterance and $d_i = \tau_{i+1} - \tau_i$ is the duration in frames of the ith phonetic segment. 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, the following "local average log-likelihood" score L can be used:

$$L = \frac{1}{N} \sum_{i=1}^{N} \frac{l_i}{d_i}, \tag{3}$$

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

The log-likelihood scores can be used for both text-dependent and text-independent scoring. However, there is no normalization against speaker variability and, as we will see in Section 5, these scores exhibit low correlation to human expert ratings.

4.2. Normalized acoustic scores

The correlation between the acoustic log-likelihood scores presented in Section 4.1 and the human expert ratings can be improved if the former are normalized using some estimate of the degree of match between the spectral characteristics of the student speaker and the training native speakers. This normalization can be achieved by using the scores of a set of context-independent phonetic models.

One approach is to use a phone-normalized score $\tilde{L}$ (Digalakis, 1992), where the log-likelihood of each phonetic segment is normalized by the log-likelihood of the context-independent phone model that better matches the observations within that segment,

$$\tilde{L} = \frac{1}{N} \sum_{i=1}^{N} \frac{l_i - L_i}{d_i}, \tag{4}$$

where $L_i = \max_q l_i(q)$ is the maximum log-likelihood score over all context-independent phones q.

The phone-normalized score may require additional computation and introduces some inhomogeneity when the segment log-likelihoods are computed using context-dependent phonetic models. In addition, it has a somewhat ad-hoc nature. A more elegant normalization method is to compute the average posterior probability for each phone (Franco et al., 1997). The motivation behind using posterior scores is that the better a student has pronounced a certain phonetic segment, the more likely this phone will be over the remaining phones when the likelihoods are being computed using native-speaker models. Posterior scores are normalized using the average, rather than the maximum, log-likelihood of each frame.

Specifically, for each frame of the ith segment that corresponds to the phone q_i, we compute the frame-based posterior probability $\mathrm{p}(q_i|x_t)$ of the phone q_i given the observation vector

$$\mathrm{p}(q_i|x_t) = \frac{\mathrm{p}(x_t|q_i)\mathrm{p}(q_i)}{\sum_{q=1}^{M} \mathrm{p}(x_t|q)\mathrm{p}(q)}, \tag{5}$$

where $\mathrm{p}(x_t|q)$ is the probability density of the current observation, using the model corresponding to phone q. The summation in the denominator is over all context-independent phones $q = 1, \ldots, M$ and $\mathrm{p}(q)$ represents the prior probability of phone q.

Similar to the local log-likelihood score, the logarithm of the frame-based posterior probability can be accumulated over all the frames of the ith segment,

$$\rho_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \log(\mathrm{p}(q_i|x_t)). \tag{6}$$

The posterior-based score for a whole sentence P is defined as the average of the individual posterior scores over the N phone segments in a sentence, normalized by their durations,

$$P = \frac{1}{N} \sum_{i=1}^{N} \frac{\rho_i}{d_i}. \tag{7}$$

Both the phone-normalized and the posterior-based scores should be less affected by changes in the spectral match due to particular speaker characteristics or acoustic channel variations. The same changes in acoustic match would similarly affect both the numerator and denominator in (7), making the score more invariant to those changes and more focused on pronunciation quality.

4.3. Segment classification scores

Pronunciation can be accessed by using a measure based on recognition error; if the recognizer is trained with native speakers, then the more the pronunciation of the test speaker resembles that of the training population, the higher the recognition accuracy should be. One approach is to use the word error rate, that is, the percentage of the words that are either misclassified, deleted or inserted by a word recognizer. If automatic and easy development of new lessons is desired, however, the word recognizer must have a large vocabulary and a very general language model. With today's state-of-the-art speech recognizers (Digalakis et al., 1996), it is not feasible to achieve good performance for nonnative speakers, especially without adaptation (Digalakis et al., 1995). Our solution to this problem is to use a phone recognizer, with a grammar at the phonetic level. If the phone recognizer is trained with native speakers, then the phone recognition accuracy can be used as a pronunciation score.

4.4. Segment duration scores

For psychological and linguistic reasons, relative phone duration should correlate well with the human expert listener's scores. The cognitive load of thinking about how to articulate can disrupt the speech flow and increase disfluency. Cross-language differences between the native language and the language being learned can also affect durations of segments. Differences in letter-to-sound rules for the orthographies of two languages may lead to insertions, deletions or substitutions of phones that will result in duration differences.

Duration scores can be obtained by measuring, from the Viterbi phonetic alignment, the duration in frames for the ith segment; then, its value must be normalized to compensate for the rate of speech (ROS) of the particular speaker. The corresponding segment duration score can be obtained by computing the log-probability of the normalized segment duration, using a discrete distribution of durations for the corresponding phone. These discrete duration distributions can be trained from alignments generated for the native training data. Hence, the segment duration score can be defined as

$$D = \frac{1}{N} \sum_{i=1}^{N} \log(\mathrm{p}(f(d_i)|q_i)), \tag{8}$$

where $f(d_i)$ is the duration normalization function and q_i is the phone that corresponds to the ith segment.

In text-dependent methods, one can normalize the duration of the ith segment by the duration d_{w_i} of the word w_i in which it appears, $f(d_i) = d_i/d_{w_i}$ (Digalakis, 1992). Since each word will have appeared many times in the native training data, discrete distributions for the word-normalized duration of the different phones can be estimated. However, to achieve text independence, we cannot use sentence, phrase or word durations to normalize phone durations. We use a measure of ROS as the normalization factor (Neumeyer et al., 1996). The simplest approach to ROS is to compute the global ROS as the average number of phones per unit of time for a given speaker. Normalized duration can be computed as

$$f(d_i) = d_i \cdot \mathrm{ROS}_s, \tag{9}$$

where ROS_s is the estimated ROS for speaker s. To compensate for phone alignment errors near silence, we can exclude phones in the context of silence from the training and testing data sets.

4.5. 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.

5. Experiments

We experimented with the VILTS corpus for the various machine scores presented in Section 4. First, a native French recognizer was trained by SRI's Decipher™ speech recognition system (Digalakis et al., 1996); the training involved 16,000 utterances from 100 native speakers of Parisian French reading newspaper text. 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 by using the Viterbi algorithm with the native French models.

The pronunciation scoring algorithms were evaluated using a test set with an average of 30 sentences per speaker 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.

Based on previous experimentation (Neumeyer et al., 1996), when obtaining the different machine scores for each sentence, in all the experiments we removed the scores of the phones in context with silence because their alignments might be inaccurate. Doing so produced a small but consistent increase in the correlation for all machine score types. To evaluate the different types of scores at the speaker level, about 30 sentence scores were averaged for each of the 100 speakers before the correlation was computed.

In Table 6 we show the correlations between the different machine and human scores computed at the sentence level (across 3000 sentences) and speaker level (across 100 speakers). Both global

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

Machine score	Correlation coefficient	
	Sentence level	Speaker level
Global log likelihood G	0.182	0.313
Local log likelihood L	0.285	0.481
Log posterior score P	0.521	0.842
Phone recognition accuracy	0.399	0.469
Segment duration score D	0.410	0.856
Syllabic timing	0.355	0.726

and local HMM likelihoods are poor predictors of pronunciation ratings. Local HMM likelihoods have slightly better correlation with human scores than the global HMM likelihood, which agrees with the perceptual argument given in Section 4.1.

Phone classification has a performance similar to that of the local log likelihood at the speaker level but seems to correlate better at the sentence level. The posterior-based score has the highest correlation at the sentence level, followed by the duration score, having a 20% lower correlation. At the speaker level the normalized duration and the log-posterior scores are comparable, having the highest correlations of all the machine scores evaluated and rendering a performance similar to that of the human raters, as we showed in Section 3.

Finally, the timing scores result in acceptable speaker-level correlations. Global ROS 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.

While at the speaker level the best machine scores reach a correlation level comparable to that of humans, the sentence-level correlations are still lower than those among humans, suggesting that further work is needed to predict pronunciation ratings when using only a single utterance. Using a slightly different development set, we calculated the speaker-level correlation between human and machine scores by using various amounts of test data. The human scores were the speaker-averaged scores of the 100 speakers, using the complete set of 50 sentences, in all cases. In this way we always correlated the machine scores with our best estimate of the speaker-level human score. To obtain the speaker-averaged machine score for variable amounts of data, we varied the number of sentences per speaker (N) from 1 to 50. For each value of N, a random subset of N sentences was chosen from the 50 speaker sentences. The speaker-averaged machine score was created by averaging the

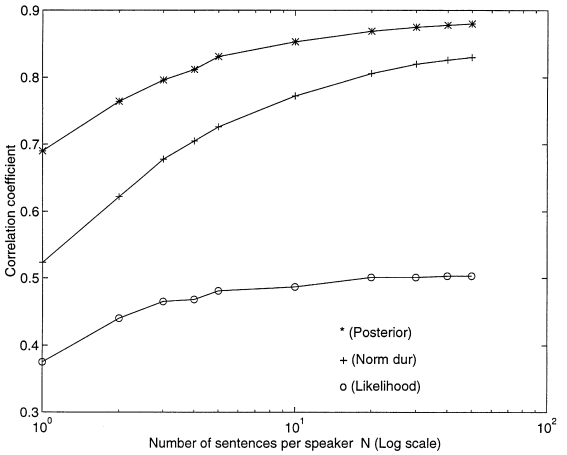

Fig. 1. Speaker level correlation for posterior, duration and likelihood scores for different numbers of sentences per speaker.

N sentence machine scores. This random experiment was repeated 40 times and the calculated correlation values for each N were averaged.

Clearly, correlations improve as the amount of test data increases. As we can see in Fig. 1 the posterior probability and the duration scores perform similarly for large amounts of data. As we use fewer sentences per speaker, the posterior score outperforms the duration score, particularly for low values of N, that is, three sentences are enough to obtain a correlation of approximately 0.8.

6. Summary and discussion

We have extensively used a pronunciation scoring paradigm for the automatic assessment of pronunciation quality by machine. In this scoring paradigm, both native and nonnative speech data are collected and a database of human-expert ratings is created to enable the development of a variety of machine scores.

The speech database design is very important, especially for text-independent pronunciation evaluation. Similarly, the reliability of the human ratings is critical, since we see pronunciation evaluation as a prediction problem, where we are trying to predict the grade a human expert would assign to a particular skill by using statistical models constructed with the speech and the expert-ratings databases.

These statistical models predict the human-expert grade by using machine scores, or predictor variables. To be useful for our purpose, these machine scores must satisfy various properties, and we have experimented with several of the machine scores.

The validity of the machine scores was tested on the VILTS corpus to evaluate the pronunciation of American speakers speaking French. We found that certain machine scores, such as the log-posterior and the normalized duration, achieve a correlation with the targeted human grades that is comparable to human-to-human correlation when a sufficient amount of speech data is available, typically a few sentences. The correlation of these machine scores with the human grades still lags the human-to-human correlation for very short amounts of speech.

The models that we used in this paper were all single-predictor models, that is, we did not try to use simultaneously different machine scores as multiple predictor variables. In the Autograder project (Digalakis, 1992) we have found that the combination of the various machine scores can improve machine-to-human correlation significantly and we will address this issue in a forthcoming paper.

Acknowledgements

Special thanks to Patti Price for her contributions and comments and Yoon Kim for his help in running some of the experiments. We gratefully acknowledge support from the US Government under the Technology Reinvestment Program (TRP). The views expressed here do not necessarily reflect those of the Government.

References

Bernstein, J., 1992. Automatic grading of English spoken by Japanese students. SRI International Internal Reports Project 2417.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M., 1990. Automatic evaluation and training in English pronunciation. In: Proceedings of the International Conference on Spoken Language Processing 1990.

Digalakis, V., 1992. Algorithm development in the autograder project. SRI International Internal Communication.

Digalakis, V., Rtischev, D., Neumeyer, L., September 1995. Speaker adaptation using constrained reestimation of Gaussian mixtures. IEEE Trans. Speech Audio Processing, 357–366.

Digalakis, V., Monaco, P., Murveit, H., 1996. Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers. IEEE Trans. Speech Audio Processing, 281–289.

Franco, H., Neumeyer, L., Kim, Y., Ronen, O., 1997. Automatic pronunciation scoring for Language instruction. In: Proceedings International Conference on Acoustics, Speech and Signal Processing 1997. Munich, pp. 1471–1474.

Greenberg, S., Baker, C., Lowe, J., 1992. Evaluating the Pronunciation of English Sentences by Japanese School Children: Validation of the Autograder Computer Model. Paper draft, September.

Kim, Y., Franco, H., Neumeyer, L., 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In: Proceedings of the European Conference on Speech Communication and Technology 1997. Rhodes, pp. 649–652.

Neumeyer, L., Franco, H., Weintraub, M., Price, P., 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In: Proceedings of the International Conference on Spoken Language Processing 1996. Philadelphia, PA, pp. 1457–1460.

Ronen, O., Neumeyer, L., Franco, H., 1997. Automatic detection of mispronunciation for Language instruction. In: Proceedings of the European Conference on Speech Communication and Technology 1997. Rhodes, pp. 645–648.

Rypa, M., 1996. VILTS: The Voice Interactive Language Training System. In: Proceedings of CALICO.