# The SRI EduSpeak™ System: Recognition and Pronunciation Scoring for Language Learning

*Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, John Butzberger, Romain Rossier and Federico Cesari*

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA
hef, victor, precoda, harry, rao, johnwb, rrossier, fico@speech.sri.com
www.speech.sri.com, www.EduSpeak.com

## Abstract

The EduSpeak™ system is a software development toolkit that enables developers of interactive language education software to use state-of-the-art speech recognition and pronunciation scoring technology. We first report results on the application of adaptation techniques to recognize both native and nonnative speech in a speaker-independent manner. We discuss our pronunciation scoring paradigm and show experimental results in the form of correlations between the pronunciation quality estimators included in the toolkit and grades given by human listeners. We review phone-level pronunciation estimation schemes and describe the phone-level mispronunciation detection functionality that we have incorporated in the toolkit. Finally, we mention some of the EduSpeak™ toolkit system features that facilitate the creation and deployment of computer-assisted language learning (CALL) applications.

## 1. Introduction

The ability to accept speech input allows developers of language learning software to move beyond traditional reading and listening comprehension activities to exercises requiring active speech production such as reading aloud, oral multiple-choice exercises, and open dialogs within limited domains. In an ideal system, the computer may provide feedback on lexical, syntactic, and acoustic aspects of language use.

Two desirable features of speech-enabled computer-based language-learning applications are the ability to recognize accented or mispronounced speech produced by language learners, and the ability to provide meaningful feedback on pronunciation quality. In this paper we report on the effort that resulted in SRI's EduSpeak™ system, a software development toolkit making these capabilities available to language-learning application developers. The EduSpeak™ recognition engine is based on the same technology as DECIPHER™, a state-of-the-art large-vocabulary speaker-independent continuous speech recognition system [1]. Its flexible acoustic modelling paradigm has allowed us to develop a high-performance real-time recognition system that is specifically adapted for CALL applications. In particular, acoustic models tailored for nonnative speech recognition, pronunciation scoring algorithms, and a number of system features were developed especially for the CALL domain.

Speaker-independent automatic speech recognition is known to degrade significantly for nonnative speakers. For instance, [2] showed error rates for nonnative speakers that were approximately double the native speaker error rate for a similar task, and [3] showed that most of the performance loss is due to acoustic mismatch; after speaker-dependent acoustic adaptation, nonnative error rates were comparable to those of native speakers.

Our goal for nonnative speech recognition was speaker-independent models that perform well for nonnative speakers without a performance degradation for native speakers. Given data from a significant number of representative nonnative speakers, approximate Bayesian adaptation methods [3] allow us to optimally combine native and nonnative training data to obtain acoustic models that perform acceptably for both types of speaker. The resulting models are text and speaker independent, so that the system does not require any user-specific registration and the models may be used for recognition tasks with arbitrary vocabulary and grammars.

The other important component of our toolkit for computer-assisted language learning is the automatic evaluation of pronunciation quality. In recent years we have developed algorithms to grade the pronunciation quality of nonnative speakers text independently; that is, these algorithms do not rely on statistics of specific words or sentences [4, 5, 6]. Text-independent scoring allows developers to use arbitrary content without the need to retrain or tune the pronunciation scoring models. Our pronunciation grading method is based on the computation of a number of text-independent scores that correlate well with human judgements of pronunciation quality. These individual scores are based on spectral, durational, and prosodic features derived from time-aligned phonetic segmentations obtained from hidden Markov model (HMM) decoding. The individual scores are nonlinearly combined to estimate the pronunciation grade that a human expert would have given. The mappings from machine scores to human grades are calibrated by using a large database of nonnative speech where individual sentences have been graded for overall pronunciation quality by human listeners [7, 8]. This approach results in pronunciation quality grades for

individual sentences or groups of sentences with grading consistency similar to that of humans.

However, an overall score is only part of the desired feedback for pronunciation training. Ideally, a CALL system should be able to diagnose specific problems in producing new sounds, and give directions for improvement. More detailed pronunciation feedback in the form of information about specific phone mispronunciations is also provided in our toolkit, allowing a CALL system to provide feedback about specific pronunciation mistakes [9].

In addition to acoustic models and pronunciation scoring algorithms, other system features were implemented to address some of the needs of software developers for CALL applications. Among these are (i) support of several development environments, (ii) small download size of both recognition engine and acoustic models to support CALL applications over the Internet, (iii) on-the-fly loading of grammars and vocabularies to support dynamic content for lessons downloaded from the Internet, and (iv) support for separate acoustic models for recognition and pronunciation scoring, respectively.

In the following sections we report some of our experimental results on nonnative recognition and pronunciation scoring as well as details of some system features.

## 2. Nonnative speech recognition for CALL

It is important to have good nonnative speech recognition performance without harming the recognition performance on native speakers such as teachers who may also interact with the system. Recognition activities with models adapted to nonnative speech can offer better nonnative performance than models trained with only native speech and thus enable more challenging tasks. It is also desirable to carry on this acoustic adaptation in such a way that good recognition performance is maintained for native speakers. To achieve both these goals, our acoustic model training approach is based on Bayesian adaptation techniques [12] that enable us to optimally combine native and nonnative training data so both types of speakers can be handled with the same models with good recognition performance. In particular we use the approximate MAP (AMAP) algorithm reported in [3] which estimates the adapted model by linearly combining the seed model and the target model sufficient statistics (usually referred to as counts) for each component density in the acoustic models.

As an example of the application of this technique, we show recognition results for a database of nonnative English. The nonnative data consisted of 12,658 sentences, read by 88 adult native speakers of Japanese with varying ability in English. The seed model was a gender-dependent, phonetically tied mixture system with 100 Gaussians per phone and was trained using the standard speaker-independent Wall Street Journal (WSJ) read speech database.

We partitioned the nonnative speech data into suitable nonnative development and adaptation sets. A native development set was also used, to monitor the performance of the adapted models on native English speech. We adapted the WSJ native models to the nonnative adaptation set, and optimized the nonnative data weight for optimal joint performance on natives and nonnatives. The nonnative development set consisted of sentences from 20 speakers balanced by gender and pronunciation skill. To find a suitable joint recognition performance we did a line search for the nonnative data weight. We observed that as the weight for the nonnative counts increases, the error rate on nonnative speakers decreases, while for native speakers it increases after an initial plateau. We chose the weight that minimizes nonnative recognition error with minimal or no degradation of native recognition.

Tables 1 and 2 show the recognition error rates for both native and nonnative speech test sets, for the baseline and the adapted models, for both male and female speakers. The recognizer used an artificial finite state grammar that is representative of language learning activities.

*Table 1: Recognition error rate in percent for male speakers in the test data.*

| Model type | Word error on nonnative speech | Word error on native speech |
|---|---|---|
| Native models | **7.49 %** | **0.58 %** |
| Mixed | **3.15 %** | **0.52 %** |

*Table 2: Recognition error rate in percent for female speakers in the test data.*

| Model type | Word error on nonnative speech | Word error on native speech |
|---|---|---|
| Native models | **6.34 %** | **0.34 %** |
| Mixed | **2.96 %** | **0.28 %** |

Using the chosen weights, relative error reduction for the nonnative speech evaluation set was 58% for male speakers and 53% for female speakers. The AMAP adaptation produced a significant improvement for the nonnative speakers without hurting performance on the native speakers. In fact, for this development set, the addition of nonnative data produced a minor increase in native performance for a range of weights.

## 3. Pronunciation scoring

The pronunciation scoring paradigm initially developed in [8, 9] and later extended to be text independent in [4, 5], uses HMMs [1] to recognize the text read by the learner and to generate phonetic segmentations of the learner's speech. With these segmentations, spectral match and prosodic scores can be derived by comparing the learner's speech to the speech of native speakers.

The generation and calibration of machine scores follows three main steps:

- Generation of a phonetic segmentation, using an HMM-based speech recognizer. The recognizer models can be trained on data from both native and nonnative speakers of the language.
- Creation of different machine pronunciation scores for the different phonetic segments by comparing features of the learner's speech to those of native speakers.
- Calibration of the scores, which includes combining several automatic measures and mapping them to estimate the judgement of human listeners as well as possible. To achieve this, it is necessary to collect pronunciation ratings by human raters of nonnative speech.

No statistics of specific sentences or words are used. Word, sentence, and speaker scores are the average of phone-level scores; consequently, the algorithms are text independent.

Our pronunciation scoring paradigm assumes that the phonetic segmentation is accurate. Therefore, the task for which pronunciation scoring is desired must be designed to ensure a high recognition rate. Reading aloud and multiple-choice exercises are examples of tasks well suited to pronunciation scoring.

The following three subsections describe machine scores previously developed that were integrated into the EduSpeak™ system.

### 3.1 Spectral match scores

We use a set of context-independent models together with the HMM phone alignment to compute an average log-posterior probability for each phone. 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 $q_i$ given the observed spectral vector $y_t$:

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

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 $q_i$. $p(y_t|q_i)$ is the probability density of the current observation using the model corresponding to the $q_i$ phone. The posterior score $\hat{\rho}_i$ for the $i$-th phone is the average of the frame-based log-posterior probability over all frames of the phone:

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

The posterior-based score for a whole sentence is defined as the average of the individual posterior scores over the N phones in the sentence:

$$\rho = \frac{1}{N} \sum_{i=1}^{N} \hat{\rho}_i$$

Since acoustic channel variations and speaker characteristics affect both the numerator and denominator of the posterior probability similarly, this score should be robust to changes in acoustic match and provide a good estimate of pronunciation quality.

### 3.2 Phone duration scores

To calculate these scores, we first obtain the duration in frames of the $i$-th phone from the Viterbi alignment. To obtain the corresponding phone duration score, the log-probability of the phone duration is computed using a duration distribution of that phone. The duration distributions have been previously trained from alignments generated for the native training data. Again, the corresponding sentence duration score is defined as the average of the phone scores over the sentence.

To take into account context or speaker dependence, we normalize phone duration by a measure of the rate of speech (ROS). ROS is the average number of phones per unit of time in a sentence or over all utterances by a single speaker.

The duration score is defined as

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

where $d_i$ is the duration of the $i$-th segment corresponding to phone $q_i$ and $f(d_i)$ is the normalizing function. Usually, $f(d_i)$ = ROS * $d_i$ .

### 3.3 Speech rate

Native speakers and advanced learners often speak faster than do beginning learners. Thus, rate of speech (ROS), alone or in combination with other scores, can be used as a predictor of the degree of nativeness. The toolkit uses the sentential speech rate, defined as the average number of phones per unit of time in a sentence.

### 3.4 Combination of scores and calibration

Two principled approaches can be devised to obtain the mappings from machine scores to human pronunciation quality ratings. The first is based on minimum mean squared error (MMSE) estimation, and the other on minimum error classification [10].

In the estimation approach, the grade that 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 by using a set of predictors, the machine scores obtained from the speech sample to be graded. Using a well-known result from probability theory, when using a MMSE criterion between the actual human grades and the predictions, the optimal predictor is the conditional expected value of the human grade given the

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

In the classification approach, each sentence is classified as belonging to one of N classes, where the classes are the discrete pronunciation grades assigned by human raters. To classify a sentence we use the optimal Bayes decision rule, which minimizes the classification error rate. The optimal predictor of the human grade is the grade with the highest posterior probability given the machine scores.

Both methods require the computation of the posterior probabilities of the human grades given the machine scores. If we do not know the mathematical form of the underlying joint probability distribution of the human and machine scores, it is necessary to resort to nonparametric methods such as neural networks, regression and classification trees, and probability distribution estimation using scalar or vector quantization [10, 7].

We have experimentally evaluated both the estimation and the classification approaches as well as the effects of the database priors in the calibration of the mappings [10]. We concluded that the priors have a strong effect on the mapping of machine scores, and that assuming equal priors produces better mappings, with higher correlation and consistency with the human grades. We also argued that the mappings obtained using the Bayes classification approach may have more desirable properties than those obtained using the estimation approach, both in being more consistent with the human scores, and in being less affected by the variability of the machine scores.

In the EduSpeak™ toolkit both the combination of machine scores and calibration are achieved using a regression tree trained on a database of nonnative speech which has been graded for pronunciation quality by human listeners and for each sentence of which we have computed all our machine scores. The training data is resampled with equal priors per grade before the classification tree is trained.

### 3.5 Experimental results for nonnative English

Previous work has shown results of our pronunciation scoring system on databases of nonnative French [4, 5, 6, 7] and Spanish [11]. In both cases the speech data was produced by native speakers of American English. In this paper we show results on a new database of nonnative English collected at SRI with the goal of applying EduSpeak™ for ESL.

The nonnative corpus includes 12,658 distinct sentences read by 88 adult native speakers of Japanese. A panel of 7 native American English speakers rated the overall pronunciation quality of a subset of 4652 sentences, using an anchored scale from 1 to 5. This set of sentences, which was rated by all the raters, was used to compute the average correlation between a rater's grade on a sentence and the grades given by all other raters. The value obtained was r=0.80, which can be considered to be an upper bound on the correlation that can be expected between machine scores and human grades. For the score correlation and the training of the mapping trees we used a subset of 4598 sentences that had both machine scores and human grades. Native acoustic models and native duration distributions were trained using the previously mentioned WSJ corpus. An additional set of native data, 422 sentences of the 1994 WSJ development set, were used to optionally extend the scoring levels with a *native* category, assigned a grade of 6.

The nonnative and native speech databases are divided into two equal-size sets with no speakers in common. We estimate the parameters of the regression tree in one set and evaluate the correlation of the predicted scores and the human grades in the other set. Then we repeat the procedure with the sets swapped and average the two correlation coefficients. We compute and evaluate the regression tree once with the additional native speakers and once without native speakers. To examine the effects of the different machine scores, we first compute the tree by using only the average log-posterior score, then add the duration score, and finally add the ROS.

We have already shown that when sufficient speech data is available for a speaker, it is possible to achieve a correlation with human grades that is comparable to the correlation between humans [5]. Our current challenge is to improve the human-machine correlation by using only the score from a single sentence.

The human-machine correlation is computed between the human grade and the conditional expected value of the regression tree's grade given the machine scores, for each sentence. Tables 3 and 4 show sentence-level human-machine correlations with mapped and combined scores, with and without the set of native speakers.

*Table 3: Human-machine correlations with mapped and combined machine scores in English, with native and nonnative speakers.*

| Machine scores | Human-machine correlation |
|---|---|
| Posterior | **0.727** |
| Posterior + duration | **0.742** |
| Posterior + duration + rate | **0.757** |

For the dataset with both native and nonnative speakers the addition of scores based on duration and rate of speech increases the correlation by 4.2%, with each individual score making an incremental contribution of 2.1%. For the dataset with nonnative data only, the addition of duration scores increases the correlation by 1.8% while the addition of speech rate has almost no effect. This result suggests that rate of speech is more useful in distinguishing native from nonnative speakers

than between different levels of nonnative pronunciation ability. Overall, the correlation is only 13% lower than the human-human correlation for a similar task

*Table 4: Human-machine correlations with mapped and combined machine scores in English, for nonnative speakers only.*

| Machine Scores | Human-machine correlation |
|---|---|
| Posterior | **0.682** |
| Posterior + duration | **0.694** |
| Posterior + duration + rate | **0.695** |

### 3.6 Phone-level mispronunciation detection

The techniques described in the previous sections allow the calculation of pronunciation quality ratings for a sentence. However, this overall score represents only part of the desired feedback for language instruction. In a classroom, a teacher can point to specific problems in producing the new sounds, and can also give instruction addressing the learner's most salient pronunciation difficulties. To provide useful automatic feedback on individual phones, we need to reliably detect whether a phone is native-like or nonnative in quality, and, ideally, to evaluate how close it is to a native phone production along different phonetic features. In previous work, posterior scores have been used to evaluate the pronunciation quality of specific phones [13] as well as to detect phone mispronunciations [14, 15]. An alternative approach [16] used HMMs with two alternative pronunciations per phone, one trained on native speech and the other on strongly nonnative speech. Mispronunciations were detected from the phone backtrace when the nonnative phone alternative was chosen.

Recently, we evaluated two mispronunciation detection schemes, one based on phone posterior scores and other based on a log-likelihood ratio (LLR) of correct and mispronounced phone models [17]. The availability of a large database [18] of phonetically transcribed nonnative Spanish allowed us to assess the performance of the two methods as well as the consistency of the transcribers [19]. For the evaluation of the mispronunciation detection, the phonetic transcriptions for each phone were mapped into two classes: native-like or not. Four phoneticians were found to disagree on which of these two classes an individual phone belonged in, in about 19.8% of cases on average. This value can be considered an approximate lower bound to the average native/mispronounced classification error. The weighted average of the minimum native/mispronounced classification error for each phone was 21.3% for the posterior-based method and 19.4% for the LLR-based method. This minimum average error can be compared with the phoneticians' pairwise disagreement above, as both take into account the prior probabilities of the data under evaluation.

Currently, the posterior-based phone-level mispronunciation detection scheme is implemented in the EduSpeak™ toolkit. For each phone we have computed a threshold corresponding to the point of equal error rate where the probability of accepting as correct a mispronounced phone (false acceptance) is equal to the probability of rejecting a phone that was correctly pronounced (false rejection). A set of 20 thresholds are defined for each phone ranging between the point of equal error rates and the point of 2% false rejections, to offer the user flexibility in the selection of an appropriate difficulty level. From the user's point of view, it seems worse to reject a correctly pronounced phone than to accept a mispronounced one. At a low rate of false rejections (and high rate of false acceptances), only the worst pronounced phones will be flagged as mispronounced. This feature enables a beginning learner to focus on a few phones most needing improvement.

## 4. System features

Several features in our toolkit facilitate the development and deployment of language-learning applications using speech recognition:

- To support CALL applications over the Internet, the acoustic model size was reduced so that the system download size for each language is under 2.5 MB including the recognition engine and acoustic models. This small size does not sacrifice recognition performance or ease of installation.
- Multiple languages are supported, including English, Spanish, and French. For most supported languages we also have at least one nonnative model set. For some languages, we have produced models targeted to children's speech, in addition to adult versions.
- Applications may be developed in C/C++, Java, and Macromedia Director. The system includes example software applications and all the documentation a developer needs to create her/his own applications.
- EduSpeak™ supports the loading of dynamic grammars and vocabulary, and rapid switching of recognition contexts.
- EduSpeak™ provides sentence-level pronunciation scoring, and sub-sentence-level pronunciation scoring by extrapolation. It also supports calibrated phone-level mispronunciation detection.
- Automatic gain control which performs on an incremental, sentence-by-sentence basis, is included.
- Different acoustic models may be used for recognition and for pronunciation scoring.
- EduSpeak™ runs on standard PCs over 200 MHz with at least 32 MB of memory.

## 5. Summary

We have presented the EduSpeak™ software development kit for development of voice-interactive language education software. We highlighted some of its features such as the availability of speaker-independent recognition models for nonnative speakers, presenting experimental results that show over 55% relative error reduction for nonnative recognition with no degradation for native speakers. We reviewed some of the fundamentals of the pronunciation scoring algorithms embedded in the system. We showed new results of pronunciation scoring on Japanese-accented English using combinations of spectral and duration scores. Correlations between human and machine scores of around 0.75 were obtained for this dataset. Our results also suggest that a rate-of-speech feature may be effective in discriminating native speakers from nonnative. We also reviewed some of the previous work on phone-level pronunciation analysis and described the phone-level mispronunciation detection scheme implemented in the system, which allows phone-level mispronunciation information to be obtained at different levels of strictness. Finally, features that facilitate system use and deployment were presented, among them a small engine and acoustic model download size, support for dynamic content, and support for several popular software authoring environments.

## References


[1] Digalakis V, Monaco P, and Murveit H (1996). Genones: Generalised Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers, *IEEE Trans. Speech and Audio Processing*, 4/4: 281-289.

[2] Byrne W, Knodt E, Khudanpur S and Bernstein J (1998). Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection and Initial Experiments in Modeling Conversational Hispanic English, *Proc. of StiLL 98*, Marholmen, Sweden, 37-40.

[3] Digalakis V and Neumeyer L (1996). Speaker Adaptation Using Combined Transformation and Bayesian Methods, *IEEE Trans. Speech and Audio Processing*, 4/4: 294-300.

[4] Neumeyer L, Franco H, Weintraub M and Price P. (1996) Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech. *Proceedings of ICSLP 96*, Philadelphia, Pennsylvania, 1457-1460.

[5] Franco H, Neumeyer L, Kim Y, and Ronen O. (1997), Automatic Pronunciation Scoring for Language Instruction. *Proceedings of ICASSP 97*, 1471-1474, Munich.

[6] Franco H, Neumeyer L, Digalakis V, and Weintraub M. (1999) Automatic Scoring of Pronunciation Quality. *Speech Communication*, 30, 83-93.

[7] Franco H, Neumeyer L, Digalakis V, and Ronen O. (1999) Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communication*, 30, 121-130.

[8] Bernstein J, Cohen M, Murveit H, Rtischev D, and Weintraub M, (1990), Automatic Evaluation and Training in English Pronunciation', *Proc. ICSLP 90*, 1185-1188, Kobe, Japan

[9] Digalakis V, (1992) Algorithm Development in the Autograder Project, SRI International Internal Communication.

[10] Franco H, Neumeyer L, (1998) Calibration of Machine Scores for Pronunciation Grading, *Proc. of ICSLP 98*, 2631-2634, Sydney, Australia .

[11] Franco H, Neumeyer L, and Bratt H (1998) Modeling Intra-Word Pauses in Pronunciation Scoring. *Proc. Speech Technology for Language Learning Wokshop (STiLL)*, 87-90, Stockholm.

[12] Gauvain J L and Lee C H, (1994), Maximum a Posteriori Estimation for Multivariable Gaussian Mixture Observations of Markov Models, *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, 291-298.

[13] Kim Y, Franco H, and Neumeyer L (1997), Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction, *Proc. of EUROSPEECH 97*, 649-652, Rhodes.

[14] Witt S and Young S (1997), Language Learning Based on Non-Native Speech Recognition, *Proc. of EUROSPEECH 97*, 633-636, Rhodes.

[15] Witt S and Young S (1998), Performance Measures for Phone-Level Pronunciation Teaching in CALL, *Proc. of the Workshop on Speech Technology in Language Learning*, 99-102, Marholmen, Sweden.

[16] Ronen O, Neumeyer L, and Franco H (1997), Automatic Detection of Mispronunciation for Language Instruction, *Proc. of EUROSPEECH* 645-648, Rhodes.

[17] Franco H, Neumeyer L, Ramos M, and Bratt H, (1999) Automatic Detection of Phone-Level Mispronunciation for Language Learning, *Proc. of Eurospeech 99*, Vol. 2, 851-854, Budapest, Hungary.

[18] Bratt H, Neumeyer L, Shriberg E, and Franco H (1998), Collection and Detailed Transcription of a Speech database for Development of Language Learning Technologies, *Proc. of ICSLP 98*, 1539-1542, Sydney, Australia.

[19] Precoda, K and Bratt, H (to appear). Perceptual Underpinnings of Automatic Pronunciation Assessment, in M.V. Holland and P. Delcloque (eds.), *Speech Technologies in Language Learning*, Swets & Zeitlinger Publishers, Lisse, The Netherlands.