Modeling Intra-Word Pauses in Pronunciation Scoring

Horacio Franco, Leonardo Neumeyer, and Harry Bratt

Speech Technology and Research Laboratory
SRI International, Menlo Park, USA
{hef,leo,harry}@speech.sri.com

Abstract

In developing computer-based systems for language learning, it is important to model some of the characteristics of the disfluent speech typical in nonnative speakers. We observed that beginning language learners often pause within words while reading. We also observed that our automatic algorithms for scoring segmental quality of pronunciation were affected by these intra-word pauses (IWPs). In this work we propose a method for modeling IWPs. As a result we are able to produce more robust segmental scores. Our experimental study also suggests that the insertion rate of IWPs could be a good predictor of fluency.

1. Introduction

In a computer-based language instruction system, the possibility of accepting speech from a student interacting with the system may allow the computer to provide feedback of the kind a language instructor would produce, such as an assessment of the quality of pronunciation or pointing to specific production problems or errors. This capability could provide valuable feedback for practicing and improving pronunciation.

Over recent years a pronunciation scoring paradigm has been developed at SRI International [1][2][3][4][5][6]. It is based on the use of hidden Markov models (HMMs) [7] to generate phonetic segmentations of the student's speech. From these segmentations, spectral match and duration scores can be derived by evaluating the closeness of a student's speech to the speech of native speakers.

Initial approaches to automatic scoring of pronunciation were based on models built for specific sentences [3]. Later, algorithms were designed to produce scores for arbitrary sentences, that is, sentences for which there is no acoustic training data [4][5]. This approach allowed more flexibility in the design of language instruction systems because new pronunciation exercises could be added without retraining the statistical models used in the scoring system.

In this paper we address some of the issues that, we believe, could arise when such a pronunciation scoring system is used in real-world situations. Often a nonnative speaker will exhibit speech disfluencies, in addition to the expected variability in pronunciation quality. In our present nonnative data collection effort we observed that one common type of disfluency was hesitation, that is, the introduction of pauses of diverse length both between and within words. While the inter-word pauses can be handled by the standard hidden Markov modeling procedure consisting of inserting optional pauses between word models, the production of intra-word pauses (IWPs) produced a strong mismatch between the spoken words and the standard word models. This mismatch is reflected as a very low pronunciation score, while in reality, the segmental pronunciation quality may have been satisfactory.

In a first step toward modeling these effects, we introduced additional pronunciation variations that allow pauses between each pair of phones within a word model. Experimental results showed that the modeling of IWPs produced word models that have a better match to the nonnative speech as well as pronunciation scores with better correlation with the human ratings, compared to the standard models with no IWP modeling. These experiments were conducted within the task of grading the pronunciation quality of native speakers of American English speaking Spanish as a second language.

2. Automatic Pronunciation Scoring

The pronunciation scoring paradigm for automatic evaluation of pronunciation consists of first eliciting read or spontaneous speech from a student, then segmenting the speech into its constituent phone units, and comparing the student's pronunciation of the phone units with that of native speakers. A pronunciation score is computed by averaging the degree of match across all the phones uttered, and this information is mapped to a scale similar to that used by human graders.

To reliably generate the phonetic segmentations for the student's speech we must know the text read by the student. We can achieve this by eliciting speech in a constrained way within the language learning activities of an automated instruction system, or simply by asking the student to read aloud prompted sentences on the computer screen. We take advantage of the constrained production by allowing only the possible alternative

word sequences in the HMMs used by the recognition system.

In our speech recognition engine [7], as in many other large-vocabulary systems, the word models are formed by a network of phone models. Therefore, the time-aligned phone sequence can be recovered using the Viterbi algorithm.

From these phonetic alignments, and statistical models obtained from the native speech, different probabilistic scores are derived for the student's speech. Since the statistical models used to perform the scoring are all based on phone units, no statistics of specific sentences or words are used; consequently, the algorithms are text independent. We have investigated several several scoring algorithms. Here, we review the two best-performing scoring algorithms, previously introduced in [4] and [5].

2.1. Log-posterior probability scores

We use a set of context-independent models along with the HMM phone alignment to compute an average posterior probability for each phone. First, for each frame belonging to a segment corresponding to the phone q_i, we compute the frame-based posterior probability $P(q_i|y_t)$ of the phone i given the observation vector y_t.

The average of the logarithm of the frame-based phone posterior probability over all the frames of the segment is defined as the posterior score for the *i*-th phone segment. The posterior-based score for a whole sentence is defined as the average of the individual posterior scores over the N phone segments in a sentence. The log-posterior score is fairly robust against changes in the spectral match due to particular speaker characteristics or acoustic channel variations [5].

2.2. Segment duration scores

To compute phone duration scores, we first measure the duration in frames for the *i*-th segment from the Viterbi alignment. Then its value is normalized to compensate for rate of speech. To obtain the corresponding phone segment duration score, the log-probability of the normalized duration is computed using a discrete distribution of durations for the corresponding phone. The discrete duration distributions have been previously trained from alignments generated for the native training data. The corresponding sentence duration score is defined as the average of the phone segment scores over the sentence

3. Modeling Intra-Word Pauses

Typically, in a large-vocabulary speech recognition system, any word to be recognized is represented in a dictionary that translates each word to a sequence of phones; each word may have multiple entries corresponding to its multiple pronunciations. For each word, these different pronunciations are compiled into an HMM, that represents the alternative pronunciations as alternative branches in the model.

Our approach to modeling intra-word pauses was to introduce additional pronunciations for each word in the vocabulary by introducing an optional pause model between every possible pair of adjacent phones in the original pronunciations for a word.

The pause model, like other phone models in the recognition engine, is formed by a sequence of three states with no jumps across states. Hence, such a model enforces a minimum duration constraint of three frames for the pauses, i.e. 30 ms given the standard frame rate of 10 ms.

In preliminary experimentation we noticed that sometimes short pauses, of a few frames, would be inserted in places where actually there was no perceptible pause in the production of the sentence. The inserted pause simply was a better match than a native phone model for that segment of the nonnative speech.

We attempted to reduce or eliminate these false pause insertions by enforcing a longer minimum duration for the pauses. We systematically modified the pause model to include more states to explore the effect of different constraints for the minimum duration of the pause models. This was implemented by concatenating multiple copies of the original pause model when a pause had to be inserted between two phones. The expanded multiple pronunciation word models were concatenated to build the necessary sentence models, also allowing the standard, optional, looped pause model between words.

4. Experimental Results

We give an overview of the speech databases used to train and evaluate the scoring models, and then we present the experiments assessing IWP modeling.

4.1. Training and calibration data

The acoustic models used to generate the phonetic alignments and produce the posterior scores were gender independent, Genonic Gaussian mixture models introduced in [7]. These models were trained using a gender-balanced database of 142 native Latin American Spanish speakers, totaling about 32,000 sentences.

For the pronunciation scoring experiments we used a database which included 206 nonnative speakers whose native language was American English. The speech material consisted of 14,000 read newspaper sentences. All the speech was recorded in standard offices with computers running, using a high-quality Sennheiser microphone.

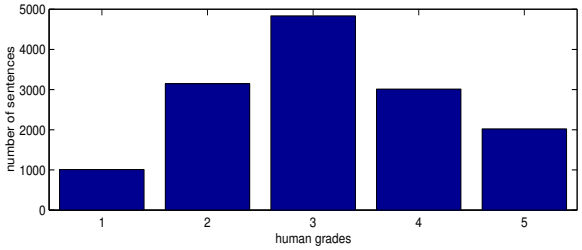

Figure 1: Histogram of human grades for the nonnative sentences.

A panel of five raters, native Spanish speakers, rated the overall pronunciation of each nonnative sentence on a scale of 1 to 5, ranging from "strongly nonnative" to "almost native". The resulting distribution of sentence grades is shown in Figure 1.

These human grades were used both to evaluate the effectiveness of the different machine scores as predictors of the pronunciation quality, and to calibrate the mappings from the machine scores to the predicted pronunciation grades. To assess the consistency of these human scores, two types of correlations were computed. At the *sentence level,* pairs of corresponding ratings for all the individual sentences were correlated. At the *speaker level*, first, the scores for all the sentences from each speaker were averaged, and then the sequence of pairs of corresponding average scores for each of the speakers was correlated. The correlation between raters was computed in a subset of 2800 sentences that were rated by all five raters. The average sentence/speaker level inter-rater correlation was r=0.68/0.91.

4.2. Scoring with IWP models

To assess the effectiveness of the IWP-based machine scores to predict the pronunciation quality, we evaluated the level of correlation between human grades and machine scores. We used posterior and duration scores computed using word models with a range of various IWP minimum duration constraints. Sentence-level human-machine correlations were computed using the 14,000 sentences from the nonnative database.

In Figure 2 we show the sentence-level human-machine correlations for both posterior and duration scores, for different values of the minimum duration pause imposed by the IWP models. The minimum duration constraint ranged from 30 to 600 ms The dotted lines correspond to the baseline system with no IWP models.

For posterior scores, we observe that there is an increase in the level of human-machine correlation when using IWP models with respect to the baseline. The correlation reaches r=0.59 when the minimum duration constraint is between 240 and 270 ms This represents a relative increase of 8.6% with respect to the baseline of r=0.54. This difference is significative at the 0.95 level.

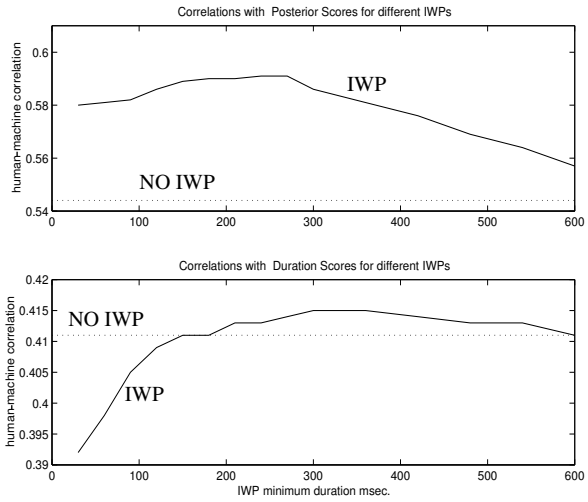

Figure 2: Human-machine correlation as a function of the minimum duration constraint for the IWP models for posterior and duration scores; the dotted lines show the correlation for the baseline models.

For duration scores, the correlation reaches a peak of r=0.42 when the minimum duration constraint is between 300 and 360 ms; the increment with respect to the baseline of 0.41 is not significant.

In Figure 3 we show the percent of inserted IWPs for each minimum duration constraint. It is interesting to note that for the best case, only 1.4% of the words have IWPs. We have not yet assessed the accuracy in detecting the actual pauses by using IWP models; nevertheless, an informal check of about 50 sentences with IWP showed that when the durational constraints were only a few frames long, many short IWPs were falsely inserted. They acted as a filler model for the segments where there was a significant acoustic mismatch between the nonnative speech and the native models. This false IWP insertion also explains the decrease in correlation observed in Figure 2 for duration scores at low values of IWP minimum duration. The false IWP insertion rate is reduced greatly by enforcing a larger minimum duration. It is worth noting that the level of human-machine correlation peaks when the minimum duration constraint is close to the average length of a phone.

For the best case of IWP modeling, corresponding to the minimum duration of 240 ms, we plotted in Figure 4 the percent of IWPs for each human grade. We observe that

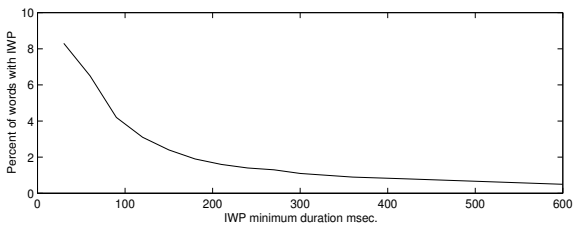

Figure 3: Percent of words with IWPs as a function of the minimum duration constraint.

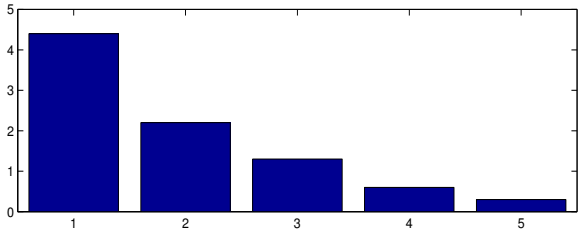

Figure 4: Percent of words with IWPs for each grade.

most of the inserted IWPs occur in the sentences with the lower pronunciation grades.

As we expected, one effect of IWP modeling was to make the machine scores more consistent. This was reflected as a significant reduction in the standard deviations of the conditional distributions of machine scores per human grade. The reductions were 27%, 29%, 19%, 10.3% and 5% for the corresponding grades 1 to 5, respectively.

A general nonlinear transformation, implemented by a neural network, can optimally map one or more machine scores to the corresponding human grades [5]. The evaluation data was divided into two halves with no common speakers, and a network was trained in each half to map the machine scores of the other half. The human-machine correlation results obtained in each half were averaged and are shown in Table 1 for both standard and the best IWP models. The correlations with the original raw scores (with no mapping) are also shown.

We observe that, for the posterior scores, the nonlinear map produced a much larger improvement on the baseline no-IWP models than in the IWP-based models. Analyzing the mappings we found that those corresponding to the no-IWP models were highly nonlinear while those corresponding to the IWP-based models were quasi-linear. The effect of the nonlinearity on the no-IWP scores was to compress the range of the mapped scores in the region of very low machine scores, which are mostly associated to the mismatched IWPs.

Table 1: Human-machine correlations with mapped and combined machine scores

Machine scores	**Human-machine correlations**	
	baseline	**IWP**
Raw posterior score	0.542	0.587
Mapped posterior score	0.586	0.593
Raw duration score	0.408	0.412
Mapped duration score	0.416	0.417
Mapped posterior + duration	0.607	0.609

The duration scores did not show significant differences between the mapped and unmapped cases. They produced a small but significant improvement in the overall correlation when combined with the posterior scores.

5. Discussion and Summary

We described the effect of modeling IWPs in a pronunciation scoring system. We showed that it leads to more consistent machine scores. Nevertheless, because IWPs appear mainly in sentences with the lowest scores, the nonlinear mapping of machine scores allowed the non-IWP-based models to reach values of human-machine correlation similar to those of the IWP-based models. We argue that IWP modeling allows evaluation of the segmental pronunciation quality independently from the level of fluency. Furthermore, the detected IWPs could be used to derive an independent machine score to evaluate the level of this particular type of disfluency.

Acknowledgments

Special thanks to María Ramos Martorell for help in running some of the experiments. We gratefully acknowledge support from the U.S. Government under the Technology Reinvestment Program.

References

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

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

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

4. L. Neumeyer, H. Franco, M. Weintraub, and P. Price, " Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech", *Proc. of ICSLP 96*, pp. 1457-1460, Philadelphia, Pennsylvania, 1996.

5. H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, " Automatic Pronunciation Scoring for Language Instruction", *Proc. Intl. Conf. on Acoust., Speech and Signal Processing 97,* pp. 1471-1474, Munich, 1997.

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

7. V. Digalakis and H. Murveit, " GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer", *Proc. of ICASSP94,* pp. I537-I540*,* 1994.