

PROSODIC FEATURES FOR AUTOMATIC TEXT-INDEPENDENT EVALUATION OF DEGREE OF NATIVENESS FOR LANGUAGE LEARNERS

Carlos Teixeira^{1,2}

Horacio Franco²

Elizabeth Shriberg²

Kristin Precoda²

Kemal Sönmez²

¹IST-UTL/INESC, Lisbon, Portugal

²SRI International, Menlo Park, CA 94025

ABSTRACT

Predicting the degree of nativeness of a student's utterance is an important issue in computer-aided language learning. This task has been addressed by many studies focusing on the segmental assessment of the speech signal. To achieve improved correlations between human and automatic nativeness scores, other aspects of speech should also be considered, such as prosody. The goal of this study is to evaluate the use of prosodic information to help predict the degree of nativeness of pronunciation, independent of the text. A supervised strategy based on human grades is used in an attempt to select promising features for this task. Preliminary results show improvements in the correlation between human and automatic scores.

1. INTRODUCTION

The goal of this study is to evaluate the use of prosodic information to help predict the degree of nativeness of pronunciation in a text-independent task. This task has been addressed by many studies focusing on the segmental assessment of the speech signal [1, 2, 3, 4]. To achieve improved correlations between human and automatic nativeness scores, other aspects of speech should also be considered, such as prosody. Recently, several studies have used suprasegmental speech information for computer-assisted foreign language learning [5]. The present work's contribution is to attempt to select promising features, using a supervised selection strategy based on human scores.

We expect prosody to carry information about the degree of nativeness of both sentences and individual words. In this study we concentrate on the effects at the word level. Our methodology was based on three steps:

1. Feature extraction. Durational and melodic information was obtained from each sentence in the form of
 - Time alignments, obtained with SRI's DECIPHERTM HMM-based speech recognition system [6].
 - Stylized pitch contours, from a model of dynamic prosodic information [7].

Potentially useful and meaningful features were derived from this information and combined with lexical information.

2. Prosodic modeling. Decision trees were used to produce the automatic nativeness scores. These trees were generated using the same procedures and parameters as in previous studies [1].
3. Combination with other knowledge sources. The prosodic features used in this work were combined with previously computed scores of the degree of nativeness — based on spectral match and timing information [2] — in order to achieve higher correlations with scores given by human listeners.

We report results obtained from the time alignments, from the pitch contours, and from their combination. These results are compared with results using other features based on segmental information [1].

2. SPEECH DATA AND HUMAN SCORING

The database contained nearly a hundred adult native Japanese speakers. The set of speakers was fairly balanced on the basis of gender and English pronunciation abilities, which ranged from beginning to advanced. Each speaker read 145 sentences taken from a pool of 12,000 different English sentences. These included sentences from news broadcasts, literature, children's literature, and simple sentences written expressly for this use.

In order to normalize some of the features, a native-speaker corpus was also required. A subset of the Wall Street Journal (WSJ) speech corpus was used for this purpose.

Each utterance from the nonnative speech corpus was graded by seven native American English speakers. The ratings were on a scale from 1 to 5, where a rating of 5 indicated very good pronunciation, and a rating of 1 indicated that the utterance had a strong foreign accent. The average correlation between the raters was computed to be 0.8 [1]. The median of the ratings from all raters was found for each utterance. These values were used as the reference human scores and served as the inputs for the supervised classification approach used in this study.

3. FEATURE DESCRIPTION

For the goals of this work, the features derived should obey three requirements: be independent of the words of each utterance to be scored; capture the essential characteristics of the different degrees of nativeness of student speech rather than being limited to any particularities of the test database; and be classified in such a way that the influence of segmental and suprasegmental information can be distinguished, to allow the investigation of advantages of using suprasegmental features in addition to the segmental ones.

In accordance with the first requirement, many of the features are averages of measurements taken over the time. The remainder resulted from unique events that were uniquely defined in each utterance, such as the maximum or minimum of a feature.

Gender was the only feature assumed to be known and the only one clearly based on specific speaker characteristics. Gender can be either provided or estimated automatically [8].

Most of the features proposed are based on durations, normalized by the rate of speech (ROS) [9], which was itself used as a feature. The phone durations used were further normalized by the average phone durations estimated from a native English corpus (WSJ).

To define features related to prosody, we estimated a time instant for the primary stress in each word. These instants were then used as references for providing text-independent information. Three definitions of the time of primary stress were computed:

- The center of the longest vowel within each word, according to segmental forced alignments.
- The center of the vowel carrying primary lexical stress.
- The instant of time of maximum F0 excursion within each word. The nearest vowel to this instant was taken to be the primary stressed vowel.

Using each of these definitions we computed three features that we refer to as the *word stress* features: the duration of the assumed primary stressed vowel, the duration between the center of this vowel and the center of the next vowel within the word, and the duration between the center of the assumed primary stressed vowel and the center of the previous vowel within the same word.

3.1. Features derived from forced alignments

The following features are average durations, computed only with the information provided by the Viterbi forced alignments. We used averages of the duration of intra-sentence pauses, the time between these pauses, and duration of words, vowels, and time between centers of vowels. A subset of the WSJ database was used to compute the

average native duration for each vowel in the phone inventory. The duration of each vowel in the utterance was normalized by the corresponding native average and used as a feature. Within each word the longest vowel was found and the word stress features were computed.

3.2. Features based on alignments and lexical stress information

The lexical primary stressed vowel of each word was located in the forced alignments. Using this vowel, the word stress features and the duration to the next lexically stressed vowel (in a following word) were computed. This last feature represents an approach to estimating rhythm. The average time difference between the maximum F0 excursion and the longest vowel in the word completed this set of lexical features. These features were averaged over all words containing lexical primary stress in the utterance.

3.3. Features based on the pitch signal

Pitch measurements were based on the output from a piecewise linear (pwl) model for the pitch contour [7]. This model also provides the probability, for each frame (10ms), of pitch halving or doubling. If the probability of pitch halving or of doubling was greater than the probability of neither halving nor doubling, the pitch value for this frame was not considered. The maximum F0 excursion over all frames t within a period of time T is computed with

$$\Delta_{f_0}(T) = \log \frac{\max_{t \in T} f_{0_{pwl}}(t)}{\min_{t \in T} f_{0_{pwl}}(t)} \quad (1)$$

The maximum F0 excursion within the utterance was taken as a feature. The maximum and the minimum values for the pitch slope were found within each utterance and used as features. Based on pitch slope, each frame was also categorized as unvoiced, rising, or falling. Using these categories as a stream of symbols, a bigram was estimated for each utterance. The corresponding relative frequencies of transitions between categories were used as features. The number of rising frames before the maximum F0 excursion, and the number of falling frames after this instant, were both used as features. The number of changes in slope per frame was considered another feature attempting to capture the pitch variation.

We also computed the average duration of rising regions and the fraction of time these occupied within the utterance. The maximum duration of consecutive rises was computed as well as the increase in pitch inside this rising region. Similar features were computed for the falling frames. The ratio of the number of pitch rises to the number of pitch falls was also computed.

3.4. Features based on alignments and pitch information

Combining the information contained in the forced aligned transcriptions with the pitch information enables us to find the instant of maximum F0 excursion within

each word and to measure time between this instant and other speech events found in the alignments. These measurements were then averaged for all the words in the utterance. This set of features included the value of the maximum F0 excursion, the time between the maximum F0 excursion and the center of the nearest vowel, the time between the maximum F0 excursion and the center of the longest vowel in the word, and the word stress features considering the maximum F0 excursion as the location of primary stress.

3.5. Measuring unique events

Most of the features previously described are averages of events that can occur several times in the utterance. These kinds of features are more reliable for a text-independent approach; however, some unique events can convey important information about the degree of nativeness of an utterance. Three types of events were considered: the two longest within-sentence pauses, the two longest words and the two longest vowels within the utterance. The durations of each of these were taken to be features. For the two longest words we also measured the word stress features associated with the three different methods for defining the instant of primary stress.

4. SCORING AND EVALUATION

Output of machine scores. Decision trees provide scores that can be evaluated by different measures of performance [3]. When the goal is to find a discrete score, as was asked of the human listeners, the highest posterior probability overall possible discrete scores (h_i) given the machine score \bar{m} can be used:

$$\tilde{h}_{opt} = \arg \max_{i \in \{1, \dots, G\}} [P(h_i | \bar{m})] \quad (2)$$

where G is the number of distinct grades.

A continuous score can also be derived. According to the minimum error criterion the optimal score is given by

$$E[h | \bar{m}] = \sum_{i=1}^G h_i \cdot P(h_i | \bar{m}). \quad (3)$$

Evaluation of machine scores. Two measures of performance were used on both discrete and continuous scores: the correlation and the error between the human and the automatic scores. This error is the average of the absolute value of the differences between the two scores. It is presented here as a percentage of the maximum error (difference between the highest and the lowest score of the scale used by the human listeners, i.e., four).

5. RESULTS AND DISCUSSION

Preliminary results are presented in Table 1 where each row describes an experiment done with a specific set of features: (a) features estimated using only the pitch signal as described in subsection 3.3, (b) features estimated

using only the forced alignments as described in subsection 3.1, (c) features combining the primary lexical stress information (subsection 3.2) used together with features used in experiment (b), (d) features combining alignments with primary lexical stress information and pitch (subsections 3.3 and 3.4), (e) features computed over unique events (subsection 3.5), (f) all features of the previous experiments (a-e) used together, (g,h) ROS and gender used as single features, and (i) all features in the previous experiments (a-h) combined. The experiments from (a) to (d) exclude the features measured on unique events (subsection 3.5).

Features	discrete scores		continuous scores	
	corr.	error	corr.	error
(a) pitch	0.211	27.2	0.232	25.8
(b) alignments	0.280	25.2	0.318	24.9
(c) +lexical	0.266	25.5	0.317	24.9
(d) +pitch	0.296	26.2	0.342	24.5
(e) unique ev.	0.355	24.2	0.388	24.0
(f) all the above	0.372	25.3	0.422	23.3
(g) ROS	0.276	29.2	0.284	25.1
(h) gender	-	30.8	-	26.5
(i) all the above	0.353	25.9	0.434	23.2

Table 1: Correlation and error (%) between human and machine scores. The machine scores were obtained using different combinations of new sets of features.

The use of features based on the forced alignments as in (b) provided a significant performance improvement when compared with the features based on the pitch signal (a) (increase of 37.1% in the correlation for continuous scores). The decrease in performance from (b) to (c) (-5.2% in the correlation of discrete scores) can be explained by the fact that (b) includes features similar to those used in (c). While in (b) those features were computed using the longest vowel in the word as a time reference, in (c) a similar feature was included which substitutes the position in time of the lexical primary stress for that reference. This latter reference is not very reliable for several reasons: nonnative speakers do not always place primary lexical stress correctly, and our procedures are not yet capable of handling alternative transcriptions. Combining pitch information with the previous features (d) produced improvements compared to (a-c) (7.7% in the correlation for continuous scores between (b) and (d)).

Features based on unique events (e) are very similar to some of the features used in (a-d). They are also generally estimated in smaller portions of the speech signal; however, they produced a relative increase of 20% in the correlation of discrete scores. When combined with the previous features (d) they provide a relative increase of 25.6% for the same scores (f).

As expected, ROS seems to be an important feature for scoring nativeness (g), provides better results than the features based on pitch (a), and gives results similar to those obtained just from the alignments (b). Gender, as a single binary feature, could not possibly provide more than two different scores (h). In fact, the decision trees

failed to grow more than one level.

When the ROS and gender were used together with all the features (i), the discrete scores perform worse (-5.0% in the correlation) while the continuous scores seem to perform better (2.7%). The gender information appears to be responsible for the different behavior of these scores, but this needs to be further investigated.

The following set of experiments (Table 2) used a set of machine scores that represent the state of the art in this area [2, 1], and are shown here as a reference and starting point for combining these scores with our new features. The posterior (j) and duration (k) scores present better performance than any other set of new features described here. As in previous studies, ROS was tested together with these features (m,n). The highest correlation with continuous scores was obtained, as expected, with the use of posterior, the ROS, and the duration score together (n). However, for discrete scores, a slightly better result was obtained using only the posterior and the ROS (m).

Features	discrete scores		continuous scores	
	corr.	error	corr.	error
(j) duration	0.394	24.0	0.439	23.2
(k) posterior	0.616	17.7	0.680	18.3
(l) +duration	0.659	17.7	0.695	17.9
(m) post+ROS	0.661	17.1	0.699	17.8
(n) +duration	0.653	17.5	0.701	17.7
(o) segmentals	0.675	16.9	0.711	17.5
(p) +lexical	0.674	16.9	0.708	17.6
(q) supraseg.	0.217	26.8	0.247	25.4
(r) all the above	0.661	17.4	0.707	17.7

Table 2: Correlation and error (%) between human and machine scores. The machine scores were obtained using different combinations of previously studied scores (posterior, duration, and ROS) and new sets of features.

The remaining experiments aim to distinguish the performance of features based on segmental information (o) from performance obtained just with the pitch signal (q). We considered posterior, duration, and ROS scores as segmental information together with all the new features that do not use pitch or lexical stress information (o). In (p) lexical-stress-based features were combined with segmental features (o). The last experiment includes all the features described in this paper (r).

Using only the segmental features (o) provided the best result. The improvements found in correlation, relative to the features used in previous studies, are 3.4% for the discrete scores and 1.4% for continuous scores. As before, combining lexical primary stress information did not improve performance (p). The use of the suprasegmental features (q) revealed an increase of 6.8% in the correlation of the continuous scores when compared with (a). This is probably related to the inclusion of the few features measured in unique events. Finally, combining these features with segmental features (r) did not lead to an improvement over using only the segmental features (o).

In summary, use of pitch information did not produce any improvement relative to previously used segmental scores with the features explored in this study. On the other hand, further improvements may be obtained from adding further specific features derived from the forced alignments.

6. ACKNOWLEDGMENTS

We express our gratitude to Colleen Richey and Harry Bratt for their help. This work was supported by DARPA Agreement DASW01-96-3-0001 and NSF STIMULATE IRI-9619921. The views expressed here do not necessarily reflect those of the Government. The first author also thanks the Luso-American Development Foundation and the Calouste Gulbenkian Foundation for their support.

7. REFERENCES

1. H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The SRI *EduSpeak*TM System: Recognition and pronunciation scoring for language learning," (to appear) *Proc. of Integrating Speech Technology in Language Learning*, 2000.
2. H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," *Speech Communication*, vol. 30, pp. 121-130, 2000.
3. H. Franco and L. Neumeyer, "Calibration of machine scores for pronunciation grading," *Proc. Int'l Conf. on Spoken Language Processing*, 1998.
4. L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, pp. 83-93, 2000.
5. R. Delmonte, "SLIM prosodic automatic tools for self-learning instruction," *Speech Communication*, vol. 30, pp. 145-166, 2000.
6. V. Digalakis and H. Murveit, "GENONES: Optimizing the degree of mixture tying in large vocabulary HMM based speech recognizer," *Proc. ICASSP 1994*, pp. 1537-1540, 1994.
7. K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," *Proc. Int'l Conf. on Spoken Language Processing*, 1998.
8. C. Teixeira and I. Trancoso, "Continuous and semi-continuous HMM for recognising non-native pronunciations," *Proc. IEEE Workshop ASR*, pp. 26-27, 1993.
9. L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic text-independent pronunciation scoring of foreign language student speech," *Proc. Int'l Conf. on Spoken Language Processing*, pp. 1457-1460, 1996.