EVALUATION OF SPEAKER'S DEGREE OF NATIVENESS USING TEXT-INDEPENDENT PROSODIC FEATURES

Carlos Teixeira[1,2], *Horacio Franco*[2], *Elizabeth Shriberg*[2], *Kristin Precoda*[2], *Kemal Sönmez*[2]

[1]IST/INESC-ID, Lisbon, Portugal
[2]SRI International, Menlo Park, CA 94025
carlos.teixeira@inesc-id.pt

Abstract

Giving feedback on the degree of nativeness of a student's speech is an important aspect of computer-aided language learning. This task has been addressed by many studies focusing on the segmental assessment of the speech signal. To better model human nativeness scores, other aspects of speech should also be considered, such as prosody. This study examines the use of prosodic information to evaluate the degree of nativeness of student pronunciation, independent of the text. Supervised strategies based on human grades are used in an attempt to select promising features for this task. Previous results obtained with non-native speakers showed improvements in the correlation between human and automatic scores. New strategies were evaluated with tests including native and non-native speakers. Specific features based on durations, namely for intra-sentence pauses, revealed potential use for further improvements.

1. Introduction

The aim of this work is to examine the use of prosodic information in evaluating the degree of nativeness of pronunciation for 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]. Recently, several studies have used suprasegmental speech information for computer-assisted foreign language learning (e.g. [5]). The present work's contribution is to attempt to select promising features, using a supervised selection strategy based on human scores of nativeness. While we expect prosody to carry information about the degree of nativeness of both sentences and individual words, in this study we concentrate on 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 DECIPHER^{TM} hidden Markov model (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.

Preliminary results with non-native speakers have shown improvements in the correlation between human and automatic scores [8]. These results are now augmented with test sets that include native speakers to provide a wider range of scores as well as a richer database for the calibration of nativeness scores.

2. Speech data and scoring

The corpus 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 addition, a subset of the Wall Street Journal (WSJ) speech corpus was selected. This allowed our system to score the higher degree of nativeness for native speakers. The training part of this subset was also used to normalize some of the features from both native and non-native corpora.

2.1. Human scoring

Each utterance from the non-native 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. A score of 6 was assigned to the utterances selected from the WSJ (native speakers). These values were used as the reference human scores and served as the inputs for the supervised classification approach used in this study.

2.2. 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:

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

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

$$E[h|\bar{m}] = \sum_{i=1}^{G} h_i \cdot P(h_i|\bar{m}). \quad (2)$$

2.3. 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., 5).

3. Feature extraction

Many of the features are averages of measurements taken over the time. The remainder resulted from 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. Most of the features proposed are based on durations, normalized by the rate of speech (ROS) [4], 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: duration of the assumed primary stressed vowel, duration between the center of this vowel and the center of the next vowel within the word, and 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, time between these pauses, and duration of words, vowels, and time between the centers of vowels. A subset of the WSJ corpus 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.

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.

3.2. Features based on the pitch signal

The maximum F0 excursion within the utterance was taken as a feature [8]. 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.

3.3. 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.

3.4. Features from 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: two longest within-sentence pauses, two longest words, and 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. Results and discussion

Previous experiments [8] performed with non-native speakers were repeated, including the subset of the native WSJ corpus. A few of the results from these experiments are represented in Table 1. These experiments aim to distinguish the performance of features based on segmental information (o2) from performance obtained just with the pitch signal (q2). We considered as segmental information (o2) the three base features (posterior, duration, and ROS scores, as proposed and evaluated in [1]) together with all the new features that do not use pitch or lexical stress information. In (p2) lexical-stress-based features were combined with segmental features (o2). The last experiment includes all the features described in this paper (r2).

	discrete scores		continuous scor.	
Features	corr.	error	corr.	error
(n2) 3 base features	0.732	14.3	0.763	14.6
(o2) segmental	0.743	13.5	0.767	14.3
(p2) + lexical	0.733	13.7	0.762	14.4
(q2) suprasegmental	0.272	23.6	0.321	22.3
(r2) all the above	0.728	14.0	0.763	14.5

Table 1: *Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

The use of the segmental features (o2) provided the best result. The improvements found in correlation, relative to the features used in previous studies (n2), are 1.5% for the discrete scores (3.4% with the non-native corpus) and only 0.5% for continuous scores (1.4% for only non-natives). As before, combining lexical primary stress information did not improve performance (p2). The use of all our suprasegmental features (q2) provides little information about the degree of nativeness. Finally, combining these features with segmental features (r2) did not lead to an improvement over using only the segmental features (o2). The results presented in Table 1 confirm previous conclusions [8]. In the following experiments we decided to follow a data-driven method for selecting a good set of features, instead of comparing results from categorical sets of features (e.g., segmental versus suprasegmental).

A first approach was based on the selection of the most successful single features in terms of continuous correlation. ROS (g2), duration (j2) and posterior (k2) scores, and average duration between intra-sentence pauses (t) have presented a continuous correlation higher than 0.4. These results are in Table 2. As in previous studies the posterior scores (k2) proved to be the more effective for the present task.

	discrete scores		continuous scor.	
Features	corr.	error	corr.	error
(g2) ROS	0.371	23.3	0.440	21.3
(j2) duration	0.445	21.0	0.511	20.2
(k2) posterior	0.700	15.8	0.730	15.6
(t) between pauses	0.407	20.7	0.427	22.2
(u) all the above	0.731	14.3	0.763	14.7

Table 2: *First feature selection approach. Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

The average duration between intra-sentence pauses (t) alone produces results comparable to previously derived features ROS and duration. The histograms for each of the given scores, of the values measured for this feature, are represented in Figure 1. It is clear from this figure that natives seldom speak continuously during a period as short as 50 to 100 ms, while non-natives do it more often as their degree of nativeness decreases. On the other hand, natives seems to be more confident about talking without any recognized pause during periods longer than 300 ms, while non-natives hardly do it for more than 250 ms.

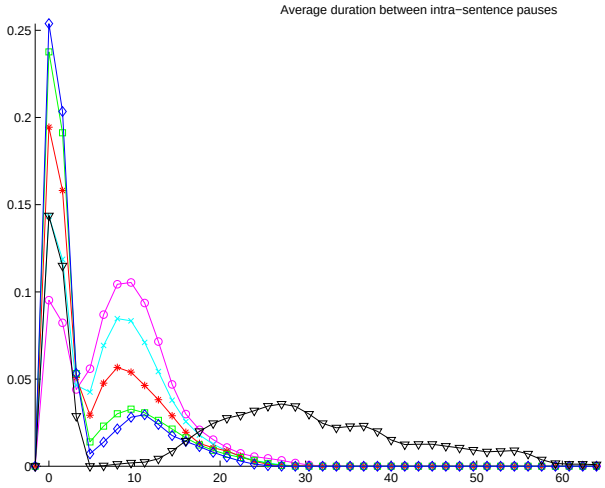

Figure 1: *Histograms of the average duration between intra-sentence pauses. Each histogram represents a single score value:* $\circ = 1, \times = 2, * = 3, \square = 4, \Diamond = 5, \nabla = 6$ *(native). The horizontal axis is in number of frames.*

Combining the single features, presenting continuous correlation higher than 0.4 (u), provides a result comparable to the use of all the features available (r2). However, this result does not show an improvement over previous studies (n2) and is not as good as the one obtained by using all the derived features based on segmentals (o2).

Table 3 represents some results obtained while following our second approach for achieving better scores while identifying additional relevant features. This approach makes use of the three base features in association with each one of the new features proposed in [8]. The results presented were selected from the experiments that have shown a continuous correlation of at least 0.765. The additional features used in these experiments were the average duration between lexically primary stressed vowels (v), average duration between the center of the longest vowel within the word and the center of the lexically primary stressed vowel (w), maximum pitch slope within the utterance (x), duration of the longest intra-sentence pause (y), duration of the second-longest intra-sentence pause (yy), longest word duration within the sentence (z), and relative frequency of the rising pitch frame followed by a falling pitch frame (zz).

The use of the longest intra-sentence pause (y) gives us an increase in the discrete and continuous correlation score of 2.5% and 0.7%, respectively, when compared with the use of the base features (n2). When compared with our previous best result (o2), these scores are only 0.9% and 0.1% better. However, the discrete correlation score of 0.75 is still the best ever found in this study. It is interesting to notice the small improvements found in experiments (x) and (zz), since the added features are based exclusively in the pitch information.

	discrete scores		continuous scor.	
Features	corr.	error	corr.	error
(n2) base features +	0.732	14.3	0.763	14.6
(v) btw_lex_stress&	0.740	13.7	0.766	14.2
(w) & max_vow	0.730	14.3	0.767	14.4
(x) max_F0_slope	0.730	14.2	0.765	14.5
(y) 1st_max_pause	0.750	13.6	0.768	14.3
(yy) 2nd_max_pause	0.739	14.1	0.766	14.5
(z) 1st_max_word	0.736	14.2	0.767	14.4
(zz) F0_rise+fall	0.735	14.2	0.766	14.5

Table 3: *Second feature selection approach. Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

Extending the principle of the first approach, allowing more than four features to be used together, we selected all features that, when used alone, provided a continuous correlation value higher than 0.10 (aa), 0.20 (ab), and 0.25 (ac) values. The more relevant results obtained for this third approach are in Table 4.

With this approach, the best continuous correlation was achieved in experiment (ab) where the selected features were: posterior and duration scores, ROS, average duration between intra-sentence pauses, duration of longest intra-sentence pause, duration of second-longest intra-sentence pause, second-longest word duration within the sentence, average duration between the center of the longest vowel within the word and the center of the lexically primary stressed vowel, maximum duration speech segment within which all frames had falling pitch, and relative frequency of a rising pitch frame followed by an unvoiced frame.

Features	discrete scores		continuous scor.	
cont. corr. $>$	corr.	error	corr.	error
(aa) 0.10	0.747	13.6	0.765	14.4
(ab) 0.20	0.740	13.7	0.769	14.3
(ac) 0.25	0.726	14.2	0.763	14.5

Table 4: *Third feature selection approach. Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

In the fourth approach, extending the principles of the second and third approaches, we selected all features that, as a result from the second approach, provided a continuous correlation value higher than 0.763 (ad) and 0.765 (ae) values. The higher correlation scores obtained are in Table 5. The second approach gave us good results, using only four features in each experiment. In the fourth approach we combined the features that provided the best results obtained with the second approach. However, this approach did not lead to a better performance than the second approach. On the contrary, the results are even slightly worse.

Features	discrete scores		continuous scor.	
cont. corr. $>$	corr.	error	corr.	error
(ad) 0.763	0.736	13.7	0.765	14.2
(ae) 0.765	0.735	13.6	0.764	14.3

Table 5: *Fourth feature selection approach. Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

In an earlier study [8], experiments made exclusively with non-native speakers did not lead to any improvement when pitch information was used in addition to the remaining proposed segmental features. This was also basically found in the experiments described in this paper, which also included a set of native speakers, apart from results (x) and (zz) in Table 3. On the other hand, improvements may be obtained from adding further specific features derived from the forced alignments. Some features based on durations — namely intra-sentence pauses — revealed potential use for improvements. We expect to continue this work in different directions. Future steps will include experiments investigating the performance of these features in discriminating between native and non-native speakers and further feature analysis and alternative supervised classification techniques.

5. 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 U.S. Government. The participation of the first author was partially supported by POSI E.U. Third Framework Programme.

6. 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 text-independent pronunciation scoring of foreign language student speech," *Proc. Int'l Conf. on Spoken Language Processing*, pp. 1457–1460, 1996.

[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, H. Franco, E. Shriberg, K. Precoda, and K. Sönmez, "Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners," *Proc. ICSLP*, 2000.