Automatic Assessment of Prosody in High-Stakes English Tests

Jian Cheng

Knowledge Technologies, Pearson
299 S. California Ave, Palo Alto, California 94306, USA
jian.cheng@pearson.com

Abstract
Prosody can be used to infer whether or not candidates fully understand a passage they are reading aloud. In this paper, we focused on automatic assessment of prosody in a read-aloud section for a high-stakes English test. A new method was proposed to handle fundamental frequency (F0) of unvoiced segments that significantly improved the predictive power of F0. The k-means clustering method was used to build canonical contour models at the word level for F0 and energy. A direct comparison between the candidate’s contours and ideal contours gave a strong prediction of the candidate’s human prosody rating. Duration information at the phoneme level was an even better predictive feature. When the contours and duration information were combined, the correlation coefficient $r = 0.80$ was obtained, which exceeded the correlation between human raters ($r = 0.75$). The results support the use of the new methods for evaluating prosody in high-stakes assessments.

Index Terms: oral reading fluency, prosody, intonation, assessment

1. Introduction
Prosody describes the patterns of stress, intonation and pausing in a spoken language. Stress is the emphasis placed on words or word segments. A working definition of intonation is the variation of pitch used when speaking in order to convey a range of meanings (questions versus statements), emotions (e.g., surprise and disbelief) or situations (sarcasm and teasing), within the confines of standard grammar and fixed word order. Intonation is reflected in the change of fundamental frequency (F0) in utterances. In addition to patterns of intonation, prosody also encompasses patterns of phoneme and pausing duration. Together, the three basic components of prosody are fundamental frequency, energy, and duration. Prosody is of interest to researchers of automated language assessments because it might be a useful tool in determining how well a candidate understands a passage that he or she is reading aloud. In this paper, we focused on automatic assessment of prosody in a read-aloud section for a high-stakes English test.

Pearson Test of English Academic (PTE Academic) [1] delivers real-life measures of test takers’ English language ability to universities, higher education institutions, government departments and professional associations requiring academic-level English. It uses automated speech scoring [2, 3] to measure the candidates’ speaking skill (pronunciation, fluency, content). The test has several different sections. In “Read aloud”, test takers are required to read aloud several passages that appear on the computer screen. The prosody used when reading these passages can be treated as a measure of reading ability.

There has been considerable research on automatic evaluation of prosody or intonation [4, 5, 6, 7, 8, 9]. Suzuki et al. [5] developed methods to evaluate the rhythm and intonation of English sentences read aloud by Japanese learners. They used F0, log power, and the first derivatives of F0 and log power as intonation features and then computed the Dynamic Time Warping (DTW) distance between native and non-native utterances. Based on the assumption that every utterance was identical to the sentence in the reading text, they used forced alignment to detect the word boundaries. Suzuki et al. then used a decision tree algorithm to cluster similar words together, assign weights to different groups of words, and generate the sentence prosody score. Because of the small number of words per utterance (between 3 and 18 words, with an average of about 6 words) [6], final correlations were low. Maier et al. [7] developed a 187-dimensional feature vector by mixing together F0, energy, and duration and then used support vector regression to predict human prosody ratings. Before being recorded, all subjects were allowed to read and practice the texts so that they could read as fluently and naturally as possible. Because the number of words in the German text used for the study was high (183 words), final correlations were high. Since F0, energy, and duration features were mixed together, it was not clear which feature contributed most to the machine-generated scores. Huang et al. [8] proposed a similar reference-independent method by incorporating more prosodic knowledge, including prosodic features based on the Fujisaki model and tempo. Since the methods used by Maier et al. [7] and Huang et al. [8] were reference independent, a completely irrelevant utterance could have produced a good score. Arias et al. [9] proposed an intonation and stress assessment system that attempted to measure the similarity between the intonation or stress curve produced by a student and a curve by a reference response. In a well controlled experimental environment with short sentences that did not include uncommon words or complicated syntactic structures, they obtained the best results by using the correlation as the similarity measure.

Unlike the data sets used in the research above, the data used here came from a high-stakes English test, which provided recordings in a real L2 assessment environment. Instead of drawing subjects from only one or two countries, the subjects used in our research came from 158 different countries and spoke 126 different native languages. Moreover, unlike other controlled environments in which researchers ensured responses were error-free so that they could use forced alignment techniques, in our less controlled environment, we could not guarantee that the responses were without error. The subjects could have made mistakes such as insertions, deletions, and/or substitutions. In addition, the test was automatically scored using automatic speech recognition (ASR). As with any ASR system, recognition performance is never 100 percent accurate. There is a certain word error rate that can possibly introduce errors to final scores. Because we were dealing with a sig-
significantly less controlled environment, we researched new approaches to predict prosody ratings.

2. Intonation and Energy Models

The F0 contour is very important for detecting intonational events such as pitch accents or boundary tones. Researchers have used many techniques to describe and label F0 contours. Some of these techniques include using features such as minimum and maximum values, onset and offset timing, and regression lines [10], or coefficients from weighted linear maximum-likelihood regression using Legendre polynomials [11], or RFC and Tilt models [12, 13]. Most of these methods focused on the contour, such as rise, fall, rise-fall, fall-rise, etc. Instead of indirectly modeling the contour, we decided to use data-based direct modeling. The assumption was that we would be able to predict the candidate’s human prosody rating from a comparison between the candidate’s contour and a set of highly-rated canonical contours.

Preprocessing We ran each response recording through speech recognition with a rule-based language model [14, 15]. The F0 and energy values were calculated at 10 ms intervals using an ESPS pitch tracker get_f0. In order to reduce the effects of both inter-speaker and intra-speaker variation, we normalized both F0 and energy values by converting them to z-scores using response specific means and variances. The recognition results were then aligned with the original passage. Any unmatched words were ignored. Each matched word within the response recording was associated with time boundary information and F0 and energy z-scores.

Modeling Because a word has voiced and unvoiced segments, traditionally when dealing with F0 contours, unvoiced segments have been handled by interpolation [9, 11, 12]. The problem with an interpolation approach is that the difference between voiced and unvoiced segments (such as durations) are lost. We proposed a special processing step by assigning the unvoiced segments a fixed low value. After that, both F0 and energy were smoothed by a mean filter with a span of 13 frames at the word level. The number of F0 and energy frames for the same word spoken by the same or different candidates could be different. Instead of using DTW to align them, all the F0 and energy frames for a word were sampled at 25 equivalent distance points that spanned the length of the word. Based on our observation, 25 sample points were enough to represent the different shapes of both F0 and energy at the word level and tended to cover more shapes than previous indirect modeling methods. A strong assumption behind using a word as a computational unit was that the F0 and energy shapes would have a very important role in predicting human prosody ratings. Another reason for focusing on a word as a unit was that the F0 and energy contours would have a very important role in predicting human prosody ratings. A typical clustering result is displayed in Figure 1. From the figure we can see that there are strong patterns for both F0 and energy contours. Clear shapes were observed for almost all the words we checked. The patterns in the first column in Figure 1 show that F0 and energy contours for the same word can vary significantly across subjects with high human prosody ratings. In fact, according to other research, there can be multiple appropriate prosody patterns for a word/passage, even for a single speaker [4, 16]. This variance within and across subjects is a strong argument in support of clustering. Many previous systems compared subjects to a very limited native model, sometimes consisting of only one or two teachers.

In order to compare the three ideal contour models (derived from clustering) to the candidate’s contour, different similarity methods were explored. The methods included Euclidean distance, correlation coefficient, autocorrelation, and DTW. After comparison, we found Euclidean distance to be most effective. We tried using correlation as the similarity measure to compute final scores, but different from Arias et al.’s results [9], Euclidean distance performed better in our setting. DTW did not help here. One reason could be that by using speech recognition, we extracted the word boundary information, and then we scaled every word interval to a standard length. It could be that these steps are similar to those used with the DTW method.

Procedure For each new response recording, we ran the speech recognition and aligned the recognized results to the read passage. For the words that had models, we compared the feature vector with the three clustering vectors individually. We chose the minimum distance as the raw score for that word and then normalized the scores for each word. The missed words were assigned the average value for the corresponding response recording. The final score for an utterance was the unweighted average. Different weighting strategies were tried, but no improvement was observed.

3. Duration Models

Duration is an important component of prosody. Even when human raters provide prosody ratings, they are often affected significantly by the correctness of duration. In the previous section, the F0 and energy models either did not specifically capture the duration information. In this section, we directly targeted duration using existing techniques. If enough samples for a phoneeme in a specific word existed, we built a unique duration model for this phoneeme in context.

Similar to Franco et al.’s method [17], duration statistics from native speakers were used to compute the log likelihood for durations of phonemes produced by candidates. Unlike the previous study, special attention was paid to the pauses produced by candidates and a separate predictor was computed based on them. The duration statistics models used for this research were built from native data for an unrelated test called the Versant English Test [2, 3]. We ran the native responses through speech recognition and accumulated duration data for every phoneeme. The statistics of the phoneeme durations of native responses were stored as non-parametric cumulative density functions (CDFs). When we computed the duration probability for a phoneeme produced by a new candidate, we checked if we had the specific CDF model with the corresponding word first. If we could not find a model for the word in context, the
phoneme duration without context was used. Given a sequence of phonemes in a recognized response \( p_i, i = 1 \ldots N \), and their corresponding durations \( D_i \), the log likelihood segmental probability for phonemes (\( \log_{\text{seg}} \text{prob} \)) was computed as:

\[
\log_{\text{seg}} \text{prob} = \frac{1}{N-2} \sum_{i=2}^{N-1} \log(Pr(D_i)),
\]

where \( Pr(D_i) \) was the probability that a native would produce phoneme \( p_i \) with the observed duration \( D_i \) in the context found. The first and last phonemes in the response were not used for the calculation of the \( \log_{\text{seg}} \text{prob} \) because durations of these phonemes as determined by the ASR were more likely to be incorrect.

The log likelihood segmental probability for inter-word silence durations, \( \text{iw}_{\log_{\text{seg}} \text{prob}} \), was calculated the same way, i.e. given a sequence of inter-word silences \( s_i, i = 1 \ldots M \), and their durations \( D_i \):

\[
\text{iw}_{\log_{\text{seg}} \text{prob}} = \frac{1}{M} \sum_{i=1}^{M} \log(Pr(D_i)),
\]

where \( Pr(D_i) \) was the probability that a native would produce inter-word silence \( s_i \) with the observed duration \( D_i \).

The duration models used here were very general. Different from the F0 and energy models that were trained using the combined development and training data, no data from the PTE Academic was used to train the duration models.

4. Experiments and Results

Experimental Data Recordings of read-aloud passages from PTE Academic’s field test data were used. The average number of words per passage was about 50. The sample rate for the recordings was 8 kHz with 8 bits (telephone band). Every response recording was rated by two different human raters. Human raters identified responses that had silence, fewer than half the expected words, or irrelevant or completely unintelligible material. These responses were excluded from our study. For valid responses, human raters rated the response recordings on a 5-point scale, with 5 representing the best prosody rating (consistent with previous research). The raters were asked to focus not only on the naturalness of prosody, but also on the degree to which the test-taker showed “comprehension of the text”. This meant that raters sometimes accepted pitch contours that were not “native-like”, as long as the phrasing and pausing were appropriate and stress was placed on the appropriate words. We also noticed that human raters gave low prosody ratings to certain native subjects because they read monotonously. Instead of using only native data as the reference [5, 6, 7], we selected responses with high prosody ratings regardless of whether the responses were from native or non-native speakers. Our goal was to cover all good prosody patterns for each passage. The reason for collecting native data was to guarantee that we had enough responses per item with high prosody ratings. All of the trained human raters had a master’s degree in language-related fields and resided in the US or UK.

For training and development purposes, 80 non-native responses and 15 native responses were collected from different
subjects for every passage. 340 valid responses from the non-native data were randomly selected as the development set, four responses per passage. The remaining valid responses made up the training set. All the final parameters used for validation were tuned on the development set, such as the fixed low value (-4.5) for unvoiced segments was tuned using this set. For the validation set, 158 subjects were randomly selected, for a total of 357 valid responses. Every response in the validation set was rated by four human raters.

**New Approach for Unvoiced Segments**

Three methods were explored to handle F0 of unvoiced segments: a) we interpolated all the unvoiced segments in the entire response recording; b) we interpolated the unvoiced intra-word segments and assigned a fixed low value to the other unvoiced segments; c) we assigned a fixed low value to all unvoiced segments. Correlations between machine-generated prosody scores and human prosody ratings are presented in Table 1. The results show that the new proposed method improved the predictive power of F0 significantly.

<table>
<thead>
<tr>
<th>Method for handling unvoiced segments</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>a). All interpolated</td>
<td>0.29</td>
</tr>
<tr>
<td>b). Only intra-word segments interpolated</td>
<td>0.54</td>
</tr>
<tr>
<td>c). All assigned a fixed low value</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Table 1: Correlations derived by using different methods to handle F0 of unvoiced segments in the development set.

**Validation Results**

Using the development set, a linear regression model was built using all four prosodic features (F0 contours, energy contours, phoneme durations and silence durations). We applied the linear regression model to the validation set and computed the correlation between the machine-generated prosody score and the average human prosody rating. Table 2 lists the correlations. For comparison, the average of the inter-rater correlations for four selected raters who rated the most in the validation set was 0.75.

<table>
<thead>
<tr>
<th>Features</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>0.67</td>
</tr>
<tr>
<td>Energy</td>
<td>0.67</td>
</tr>
<tr>
<td>F0 + Energy</td>
<td>0.75</td>
</tr>
<tr>
<td>iw_logseg_prob</td>
<td>0.54</td>
</tr>
<tr>
<td>logseg_prob</td>
<td>0.76</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 2: Correlations using different features.

From the table we can see, logseg_prob based on the log likelihood of the segmental duration probability for phonemes was the best individual predictor. The F0 and energy features had strong predictive power also. A simple average of F0 and energy improved the correlation significantly. The final combined model by using linear regression produced a correlation coefficient (0.80) that was better than the human inter-rater correlation (0.75).

**5. Conclusions**

Prosody consists of three basic components: fundamental frequency, energy, and duration. We used “Read aloud” recordings as assessment materials and explored different methods for predicting human prosody ratings in a real L2 high-stakes assessment environment. A new method was proposed to handle fundamental frequency of unvoiced segments that significantly improved the predictive power of F0. We used the k-means clustering method to build canonical contour models at the word level for F0 and energy, which provided strong predictors of prosody ratings. Combined with duration information at the phoneme level, we built a linear regression model that produced machine-generated prosody scores that correlated highly with human prosody ratings ($r = 0.80$). This correlation coefficient was even better than the correlation between human raters (0.75). Our experimental results showed that while F0 and energy contours were strong predictors of prosody ratings, duration information was the best predictive feature. The results support the use of the new methods for evaluating prosody in high-stakes assessments.

**6. References**


