Word-level F0 Modeling in the Automated Assessment of Non-native Read Speech

Xinhao Wang[1], *Keelan Evanini*[2], *Su-Youn Yoon*[2]

Educational Testing Service
[1]90 New Montgomery St #1500, San Francisco, CA 94105, USA
[2]660 Rosedale Road, Princeton, NJ 08541, USA
xwang002@ets.org, kevanini@ets.org, syoon@ets.org

Abstract

This study investigates methods for automatically evaluating the appropriateness of F0 contours in the task of automated assessment of non-native read aloud speech. The F0 contour of a test taker's spoken response is represented as a fixed-dimension vector with a word-level F0 value corresponding to each word in the prompt text. This vector is then correlated with gold standard vectors extracted from native speaker responses. Three different measures are used to describe the F0 contour within a word, including the mean of the F0 values, the difference between the mean values for each word and its neighboring words, and polynomial regression parameters. Additionally, features are developed based on a human expert's annotations, in which different types of words in a reading passage are identified as prosodically more important than others. Experimental results demonstrate the effectiveness of applying the proposed features to the automated prediction of intonation and stress scores for non-native read aloud speech.

Index Terms: word-level F0 modeling, automated speech assessment, read speech

1. Introduction

Read aloud test questions are a frequent and effective way to evaluate non-native speaking proficiency. In this task type, test takers are asked to read a text out loud (typically ranging from a sentence to a paragraph in length), and their speech is generally assessed on the dimensions of pronunciation, fluency, reading accuracy, and prosody. While the first three dimensions have received the most attention in the past, several recent research efforts have also been made to evaluate the prosodic naturalness of read speech in both computer-assisted language learning tools and automated assessment systems [1][2][3][4]. Since the F0 contour is one of the most important characteristics of prosody, and since it has a very important role in predicting human intonation ratings [5], this study focuses on F0 contour modeling for automated speech assessment.

Some previous research has evaluated the expressiveness of read speech produced by both native speakers and non-native speakers using prosodic contours. Schwanenflugel et al. [6] analyzed children's oral readings and adults' narrations on the same text, and assessed a child's oral reading fluency by correlating each child's speech to a profile of expressive reading generated from adults' speech. [4] further extended this approach and introduced a different set of word-level features related to word latency, duration, and intensity. Similar to these approaches, this study investigates the use of template models built from native read speech for evaluating a non-native speaker's F0 contour produced for the same read aloud passage. These template models are built based on the insight that fluent non-native speakers tend to use F0 contours that resemble the prosody of native speakers' readings of the same text.

The process of building the native model can be done in either a text-independent or text-dependent manner. In [7], a text-independent method was used, in which the speech was segmented into voiced and unvoiced sections, and a 187-dimensional feature vector consisting of F0 and energy measurements was extracted from each voiced segment. In contrast, most research uses a text-dependent method [2][3][4][6], in which the prompt texts are automatically aligned with test takers' speech, and then different metrics can be extracted from the F0 and energy contours over the time span of certain linguistic units, such as phonemes, syllables, and words [2][4]. Furthermore, [8] and [9] indicate that native speakers tend to assign different levels of importance to the words appearing in a reading passage when they are evaluating a non-native speaker's prosody. To accommodate this, they introduced a word importance factor; under this approach, a decision tree algorithm was used to automatically cluster the words in the passage and then weighting factors were assigned to each cluster.

Motivated by this finding, the current study employed a human expert who was experienced in spoken language assessment and human rater training to provide guidance about which parts of a reading passage containing multiple sentences the human raters should pay more attention to when rating a non-native speaker's intonation in read speech. Based on these analyses, the F0 contour shapes at several portions of the passage in addition to the sentence and phrase boundaries were determined to play an important role in the assessment of non-native prosody, including sentences with a relatively more complex syntactic structure, words that are expected to be emphasized, transition words, (e.g., *however*, *also*, *additionally*), and lists of items. Therefore, we obtained a multi-level annotation from this human expert for a random read aloud passage from a global assessment of non-native speech, as described in Section 2. In this study, we focused on non-native speakers from two countries (China and India), and further built native-speaker template models based on words from each layer of the annotation to be used for the automated assessment of non-native F0 contours.

The remainder of this paper is organized as follows: Section 2 describes the native and non-native speech corpora that are used in the study and introduces the expert annotations that were obtained; Section 3 presents the methodology we used to build the native-speaker template models; Section 4 shows the experimental results; and Section 5 summarized the main contributions of the study and indicates directions for future work.

Copyright © 2015 ISCA

2. Task and data

2.1. Task

The read aloud task in this study is designed to determine whether a non-native speaker can produce intelligible English to native or proficient non-native speakers. Test takers are presented with a passage on the screen; they then have 45 seconds to prepare and 45 seconds to read the passage out loud. The prompt text includes multiple sentences and the number of words within each passage ranges from around 40-60. Based on the test specifications, each read aloud passage should contain at least 1 complex sentence, a list of items, and at least one transition word (e.g., *however*, *also*, *additionally*). In addition, the content of the passage is generally related to passages that would be read aloud in a public setting, such as announcements, advertisements, introductions, etc. One human expert familiar with the test design and the human rating process for this task was invited to analyze one random prompt, and annotated this passage on several linguistic aspects related to the F0 contours as follows:

- <complex>*Welcome* to the *Metropolitan Job Fair*, | where *many full-time positions* are *featured*! </complex>|
- <transition>In *addition*, </transition>| *several part-time jobs* are *available* | for those who *need* a *flexible schedule*. |
- *Representatives* are *here* from *many industries*, | *including* <list>*tourism*, | *manufacturing*, | and *health care*. </list>|
- *Feel free* to *take* a *look around*, | and *ask* if you *have* any *questions*. |

This passage contains 4 sentences and 52 words. The first sentence is identified as a complex sentence due to the presence of a subordinate clause; also, one transition (*in addition*) and one list (*tourism, manufacturing, and health care*) were identified; all of the words that are expected to be emphasized are labeled in italics. Finally, boundaries of prosodic phrases are annotated mostly at the ends of sentence or at commas, with only one exception appearing in the second sentence. In Section 3, we will describe in detail how these annotations can be used in our word-level native models.

2.2. Data

In this work, we focus only on the annotated prompt text described in Section 2.1. Read aloud responses for this text were elicited from both native and non-native speakers. The native speech corpus consists of 82 spoken responses from speakers representing all major North American dialect regions; this corpus is used for building the native speaker models. The non-native speech corpus consists of read aloud responses drawn from the pilot administration of a global English proficiency assessment for adult learners of English; the participants in this study are from China (all participants from China had Mandarin as their first language (L1))and India (representing a range of L1 backgrounds). Human experts were then recruited to rate the non-native speech; instead of rating the test takers' overall reading proficiency, raters provided scores analytically on the following two dimensions: 1) intonation and stress and 2) pronunciation. Both analytic scores used a 3-point scale. For example, for the intonation and stress rating, score 3 (high-level) indicates that the speaker's use of emphases, pauses, and rising and falling pitch is appropriate to the text; score 2 (medium-level) indicates that the speaker's use of emphases, pauses, and rising and falling pitch is generally appropriate to the text, though the response includes some lapses and/or moderate L1 influence; and score 1 (low-level) indicates speaker's use of emphases, pauses, and rising and falling pitch is not appropriate and the response and includes significant L1 influence. In addition to these 3-level scores, human raters also provided a score of 0 if no response was provided or the response was completely unrelated to the prompt text; these responses were excluded from our experiments. In this work, the intonation and stress analytic scores are used for evaluation in the experiments described in Section 4.

	China	India
# Responses	202	230
# Double-scored Responses	156	181
κ	0.524	0.419
r	0.529	0.42

Table 1: Experimental responses from non-native speakers from China and India. Pearson correlation coefficients (r) and quadratic weighted Kappa values (κ) on double-scored responses are calculated to evaluate the inter-rater agreement.

2.3. Preprocessing

F0 measurements for each response were extracted using the Auto-Correlation method from Praat [10]. As the non-native speech corpus was drawn from an English proficiency assessment where multiple read responses were elicited from the same speaker, we used speaker-level z-scores to normalize the raw F0 values. We also experimented with using both interpolated F0 contours and response-level normalization, and obtained similar results.

In addition, a state-of-art speech recognizer was used to perform forced alignment between this prompt text and the read aloud responses both for native and non-native speech. We excluded responses from the experimental data if the recognizer failed to align all the words in the prompt text to the test taker's speech (3.6%). We calculated the word error rates (WER) between the prompt text and the human transcriptions of the excluded responses, where the overall WER is 28.8%, with deletion errors being the most frequent (19.7%). It is reasonable to assume that evaluating the appropriateness of intonation will become very unreliable when test takers make too many reading errors on a prompt text. Therefore, we decided to use the prompt for forced alignment and exclude responses which cannot be successfully aligned to all words in the prompt. The final experimental speech corpus includes 82 responses from native speakers, as well as 202 and 230 responses from non-native speakers from China and India, respectively. As shown in Table 1, approximately 78% of the non-native responses were double scored by two human raters, and the inter-rater agreements on intonation analytic scores are measured with both Pearson correlation coefficients (r) and quadratic weighted Kappa values (κ). The table shows that the inter-rater agreement on responses from speakers from China is much higher than that on responses from speakers from India, with κ values of 0.524 and 0.419, respectively.

Copyright © 2015 ISCA

3. Methodology

3.1. Word-level F0 Modeling

In read speech, the number of frames aligned within each word can vary substantially between different renderings by the same or different speakers. Therefore, there are several different ways to describe the F0 contour within a word [2][8]. For example, the Dynamic Time Warping (DTW) algorithm has been used to find the optimal correspondence between responses, and then the distance between corresponding frames was summed up with weights to evaluate the F0 contour of each word [8]. In another approach, Cheng sampled the F0 and energy frames at 25 equivalent distance points within a word, and then used the Euclidean distance to compare the ideal contours and test contours [2].

We first adopt the method in [4] to average F0 values over the time interval of each word. Accordingly, the pitch contour of a spoken response was represented as a sequence of F0 mean values, one for each word in the prompt text. This results in a vector of 52 F0 mean values, corresponding to the 52 words in the prompt text used in this study. In addition, we also calculated the difference between the F0 mean values from each word and its preceding and succeeding words, producing a 102-dimensional vector (since no difference values can be extracted between the first word and its preceding word as well as between the last word and its succeeding word). Furthermore, we applied first order polynomial regression to the F0 contour of each word and extracted the slope to represent the shape of the F0 contour. Higher-order linear fittings were also investigated, but no improvement was obtained.

3.2. Native Model Building

In the work of [2, 8], template models were built for each individual word. Given a test utterance, a feature vector was extracted for each word in the prompt text, and then distance metrics, such as the Euclidean distance or DTW distance, were applied to obtain a score on that word. Finally, the overall score for a spoken response was obtained by averaging word-level scores either with equal weights or unequal weights related to word importance. In the current approach, however, we represent each word with a mean F0 value (*Mean*), the difference of the mean values between words and their neighboring words (*MeanDiff*), and a linear fitting parameter (*Fit*). Using each of these metrics, a fixed-dimension vector was extracted to represent the pitch contour of a response. Afterwards, a canonical native template can be obtained by averaging all the native vectors (*average* model), which is then used to calculate the Pearson correlation with each test vector. These correlation coefficients, taken as features, are applied to predict the naturalness of a test taker's intonation.

There is a generally acknowledged fact that multiple appropriate prosody patterns exist for the same passage, and that they can vary greatly even for the same native speaker. Therefore, we experimented with two additional methods of obtaining template native models. First, we correlate the vector of a test response against each vector from each native speech and obtain the maximum correlation coefficient as a feature (*1-best* model). Additionally, we perform *k*-means clustering on the 82 native vectors and average the vectors within each cluster as a native template (*k-means* model). Given a test vector, we correlate it with averaged clustering templates and select the maximum correlation as a feature.

Country	Mean	MeanDiff	Fit
China	0.337	0.351	0.186
India	0.363	0.357	0.285

Table 2: Comparison of feature correlations when three different measurements are used for representing F0 contours within words.

3.3. Human Prosodic Annotation

The models described above are built by treating every word in the prompt text equally. However, as pointed out by human experts, human raters tend to treat the words in a prompt differently while rating read speech. Therefore, we explore the possibility of building different models according to each kind of human annotation. An intuitive way is to select words based on annotations before obtaining the representative vector of an F0 contour within read speech. Here, taking the annotation of potentially emphasized words as an example, we use the F0 mean value to represent the pitch contour within each word, and then extract a 33-dimensional vector of F0 mean values corresponding to the set of emphasized words in order to represent the whole response. In addition, it is worth noting that, except for the two words *here* and *many*, there is substantial overlap between the manually identified words and content words as defined by the Part-of-Speech (POS) tags. Therefore we also try to make the word selection based on function or content words that can be automatically labeled by POS tags. In total, 9 different kinds of vectors were extracted, each of which is based on one kind of annotated word list, including a list with all words appearing in a prompt text (word), a list with function words (function), a list with content words (content), a list with potentially emphasized words (emphasis), a list with words appearing in the complex sentence (complex), a list of words within the item list (list), a list of transition words (transition), a list of words at the prosodic boundaries (prosodic boundary), and a list of words at sentence boundaries (sentence boundary).

4. Experimental Results

4.1. Native Models

We first compare the three different measures used to represent F0 contours within words, i.e., *Mean*, *MeanDiff*, and *Fit*. The native model is computed by averaging all native vectors, and then the correlation between the averaged native model and a non-native vector is used as an intonation feature. This feature is evaluated by calculating its correlation with the "intonation and stress" score from the first human rater, which will be used across all experiments described in this section. As shown in Table 2, the MeanDiff method performs best on responses from the China data set, but on responses from the India data set the Mean method is the best. As these three measurements describe the word-level F0 contour from different aspects, they were all included in the following experiments.

Given the wide range of variation exhibited across native speakers, it is important to consider how many native responses are required to make the model robust enough. Thus, another experiment was conducted to examine the effect of the number of native speaker responses contained in the model. There are 82 responses in the native speech corpus, and the native speaker model is computed by taking the mean of the vectors across N native speakers. We extracted random subsets of these native speakers to compute the means, with N ranging from 1 to

Copyright © 2015 ISCA

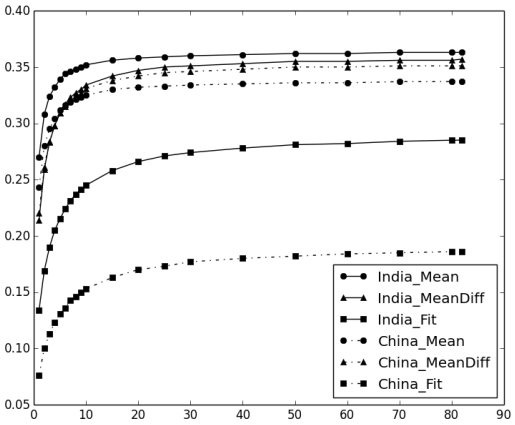

Figure 1: Feature correlations when different numbers of speakers are used in native speaker model building.

82. For each N, the experiment was run 10,000 times, and the resulting 10,000 correlations were averaged to show the performance of a mean model with N speakers. As shown in Figure 1, the feature performance fluctuates substantially when fewer than 10-20 speakers (depending on different features) are used in the model. With the addition of more native speakers beyond around 30, the improvement of the features is limited; however, the following experiments described in this section still use models based on all 82 native speakers for a thorough comparison.

Furthermore, as proposed in Section 3.2, three types of models were used to build gold standard models, i.e., *average*, *1-best*, and *k-means*. As shown in Table 3, the *average* model performs better than the *1-best* model; additionally, the *k-means* clustering method outperforms the other two methods, with the exception of the Fit F0 measure on the speakers from the India data set. But the above best performance with *k-means* was obtained when we experimented with different number of clusters, i.e., *k* value, and selected the best performing value in each individual case. It turns out that the optimal *k* value varies widely across different conditions (2-16), especially on the speech from the China dataset. Therefore, taking the feasibility of practical development into consideration (since the optimal number of clusters for a given configuration is not known *a priori*), we decided to adopt the simpler *average* native speaker model in subsequent experiments.

Country	F0 Measure	Native Models		
		Average	1-best	*k*-means
China	Mean	0.337	0.335	0.385
	MeanDiff	0.351	0.328	0.38
	Fit	0.186	0.184	0.293
India	Mean	0.363	0.344	0.373
	MeanDiff	0.357	0.295	0.37
	Fit	0.285	0.184	0.271

Table 3: Comparison of feature correlations when different methods are used to build gold standard models.

4.2. Word Selection

Different features were extracted when different lists of words based on human annotations were applied for F0 contour computation. Table 4 shows the Pearson correlation coefficients between each kind of feature and human intonation scores. As expected, on responses from both the China and India sets, the emphasis feature is superior to other features, and comparable performance is achieved between content and emphasis features. Moreover, features based on all words, phrase, and sentence also achieve relatively good correlations. As there is only one transition in this prompt text (*in addition*), the correlation for this unit is not robust; therefore, only the MeanDiff measurement is used for the transition feature. In addition, features using Fit to represent F0 contours generally perform worse than the other two measures, and most measures perform better on data from the India set than on data from the China set. Considering that the three measures (Mean, MeanDiff and Fit) can describe different aspects of the F0 contour, all of them were applied in the following experiments for the automatic prediction of intonation scores.

Country	Word Selection	F0 Measures		
		Mean	MeanDiff	Fit
China	word	0.337	0.351	0.186
	function	0.138	0.186	0.128
	content	0.34	0.346	0.153
	emphasis	0.348	0.353	0.286
	complex	0.1	0.179	0.138
	list	0.226	0.226	-0.084
	transition	-	0.105	-
	phrase	0.315	0.289	0.163
	sentence	0.261	0.255	0.112
India	word	0.363	0.357	0.285
	function	0.2	0.234	0.23
	content	0.359	0.347	0.185
	emphasis	0.366	0.356	0.299
	complex	0.247	0.268	0.277
	list	0.144	0.25	0.216
	transition	-	0.159	-
	phrase	0.335	0.341	0.299
	sentence	0.27	0.282	0.275

Table 4: Comparison of feature correlations when different lists of words are used for features extraction

4.3. Automatic Prediction of Intonation Scores

Based on the features shown in Table 4, we experimented with building a scoring model to automatically predict the intonation and stress scores from the first human rater. The linear regression algorithm from WEKA [11] was adopted to build the scoring model, and 10-fold cross validation was separately performed on both data sets from non-native speakers with different L1 backgrounds from China and India. All proposed features on words from difference levels of annotations are included in the scoring model building. Table 5 presents the results of this experiment, using the Pearson correlation coefficient as the evaluation metric; the inter-rater agreement is also included in this table for comparison. Although there is a distinct difference between human agreements for the responses from the China and India sets (with correlations of 0.529 and 0.42, respectively), the automatic models obtain very consis-

Copyright © 2015 ISCA

Country	Human Rater 2	Automatic Models
China	0.529	0.365
India	0.42	0.368

Table 5: Pearson correlation coefficients between automatically predicted scores and scores from human rater 1. The correlation between scores from human rater 1 and human rater 2 is also listed for comparison.

tent results on two different data sets. Especially on the India data set, the automatic models based on proposed features achieved a promising correlation of 0.368, compared with a correlation of 0.42 between two human raters. The wider discrepancy between human-human and human-machine agreement on the China set in comparison to the India set warrants further investigation; however, this question is out of the scope of this study, since it is likely due to additional features beyond pitch that human raters attune to in the process of scoring non-native stress and intonation.

5. Conclusion

This paper proposes various features for automatically assessing the intonation of non-native read speech by modeling word-level F0 contours. These features are extracted by correlating each test response with gold standard models built from native speech. We experiment with different methods of representing F0 contours within words, and compare different methods of building the gold standard models. Furthermore, we demonstrate an effective way to extract various features with experts' knowledge. Finally, all proposed features on words from different levels of human annotations are included to automatically predict the analytic intonation scores, and promising correlations between human and automatic scores are obtained, especially for responses from non-native speakers from China. Since this study focuses on a single prompt text, future work will examine the robustness of the proposed methods across multiple prompts. In addition, we will further examine the proposed word-level F0 features by combining them with other effective prosodic features, such as in [1], in the task of automatic prediction of intonation and stress scores.

6. References

[1] K. Zechner, X. Xi, and L. Chen, "Evaluating prosodic features for automated scoring of non-native read speech," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[2] J. Cheng, "Automatic assessment of prosody in high-stakes English tests," in *Proceedings of Interspeech*, 2011.

[3] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental English through parroting," in *Proceedings of Speech Prosody*, 2010.

[4] M. Duong, J. Mostow, and S. Sitaram, "Two methods for assessing oral reading prosody," *ACM Transactions on Speech and Language Processing*, vol. 7, no. 4, pp. 1–22, 2011.

[5] J. Tepperman and S. Narayanan, "Better nonnative intonation scores through prosodic theory," in *Proceedings of Interspeech*, 2008.

[6] P. J. Schwanenflugel, A. M. Hamilton, J. M. Wisenbaker, M. R. Kuhn, and S. A. Stahl, "Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers," *Journal of Educational Psychology*, vol. 96, no. 1, pp. 119–129, 2004.

[7] A. K. Maier, F. Hönig, V. Zeißler, A. Batliner, E. Körner, N. Yamanaka, P. Ackermann, and E. Nöth, "A language-independent feature set for the automatic evaluation of prosody," in *Proceedings of Interspeech*, 2009.

[8] M. Suzuki, T. Konno, A. Ito, and S. Makino, "Automatic evaluation system of English prosody based on word importance factor," *Journal of Systemics, Cybernetics and Informatics*, vol. 6, pp. 83–90, 2008.

[9] A. Ito, T. Konno, M. Ito, and S. Makino, "Evaluation of English intonation based on combination of multiple evaluation scores," in *Proceedings of Interspeech*, 2009.

[10] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2001.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, 2009.

Copyright © 2015 ISCA