Using F0 Contours to Assess Nativeness in a Sentence Repeat Task

Min Ma[1], *Keelan Evanini*[2], *Anastassia Loukina*[2], *Xinhao Wang*[2], *Klaus Zechner*[2]

[1]Graduate Center, The City University of New York
[2]Educational Testing Service, Princeton, NJ

mma@gradcenter.cuny.edu, {kevanini,aloukina,xwang002,kzechner}@ets.org

Abstract

In this paper, we conduct experiments using F0 contour features to assess the nativeness of responses provided by speakers from India and China to a Sentence Repeat task in an assessment of English speaking proficiency for non-native speakers. The results show that the coefficients from polynomial models of the pitch contours help distinguish between native and non-native speakers, especially among females. We find that the F0 contour can be represented adequately by using only basic statistical variables and the first three orders of polynomial coefficients. In addition, the most important features for classification are presented for each group of speakers. Finally, we discuss the differences among the gender-specific groups of the speakers.

Index Terms: F0 contour, prosody features, nativeness assessment

1. Introduction

The assessment of non-native prosody, in particular, F0 contours, is an important component of computer-based language learning and assessment tools, since deviations in prosody can often result in speech that is difficult to understand [1]. This is relevant to several different types of applications, such as automated speech assessment systems, reading tutors and computer-assisted language learning tools. Because it is often impractical for English teachers to provide individual, real-time feedback to each language learner, there is a need for a reliable automatic prosodic assessment system to provide diagnostic and corrective feedback for non-native speakers. However, prosody assessment is a difficult task due to the wide range of acceptable variation among native speakers, depending on context and individual characteristics. One successful approach has been to compare a non-native speaker's F0 contours to native speaker F0 contours, and this has been shown to provide a strong prediction of the non-native speaker's prosodic proficiency as rated by humans [2].

In this study, we apply such an approach to non-native prosody evaluation based on native speaker models. Previous studies that have used this approach have typically employed text-dependent F0 contour models by comparing non-native contours with native models for the same stimulus materials. However, such a text-dependent approach requires the existence of native models for each text and can be impractical in a large-scale language assessment where different speakers respond to many different items. Therefore, in this work we compare this traditional text-dependent approach with a text-independent approach in order to determine to what extent the features can generalize to new situations.

In addition, the F0 assessment system can be designed as either gender-dependent or gender-independent. Many previous studies of automated prosodic evaluation, such as [3], have taken a gender-independent approach, since it can be generalized more easily. However, some previous research has shown that male and female speakers have different prosodic tendencies; for example, [4] demonstrated that the distribution of pitch accents differs across genders. Therefore, we also compare the performance of generic models that combine data from both genders with gender-dependent models in this study.

There are several challenges presented in attempting to develop a system for automatically evaluating the nativeness of F0 contours in the context of a large-scale global assessment of non-native English speaking proficiency including the availability of native speaker models and the types of comparisons that can be made. First of all, obtaining a sufficient number of responses from native speakers to train native models for each prompt may be impractical or even impossible, since many responses are unscripted. Second, the low accuracy of ASR for low-proficiency non-native speakers introduces substantial noise into the computation of other features which require information about the content of the response [5]. Therefore, rather than using an abstract representation of F0 contours based on a linguistic analysis of the utterance, we focus on linguistically-naïve methods which do not rely on ASR output and compare global sentence-level properties of native and non-native intonation. We compare both F0 features based on polynomial fitting and simple descriptive statistics of the F0 measurements from each utterance.

The remainder of this paper is organized as follows: Section 2 reviews prior work on analyzing F0 contours and automated assessment of non-native prosody that is relevant to the current study; then, Section 3 presents the characteristics of the native and non-native speech corpora that were used in the study; Section 4 describes the F0 contour features and the experimental design; Section 5 presents the nativeness classification results using text-independent vs. text-dependent models and gender-independent vs. gender-dependent models as well as an analysis of the most important features; finally, Section 6 provides a discussion of the results and indicates directions for future research.

2. Previous Work

Several related studies have used F0 contours to automatically assess the oral reading proficiency of a non-native speaker. For example, [6] built canonical contour models of F0 and compared between non-native speakers with the native speakers. They modeled the contours at the word level and then computed prosody scores based on a combination of the contour features, energy features, and duration features and reported a correlation of 0.80 between these scores and human ratings.

Through an autocorrelation-based maximum posteriori approach, [7] took advantage of pitch floor and ceiling values to capture aspects of pronunciation quality not seen at the segment-level and reached a maximum accuracy of 89.8% in a classification task distinguishing between native and non-native speakers. [8] also employed a method for assessing non-native F0 contours based on comparisons with native speaker models, representing the contours in the log domain with normalization and smoothing applied. In their text- and language-independent system, the F0 contour in a non-native test utterance was compared to a reference F0 contour produced by an expert native speaker frame-by-frame by using DTW alignment. This method achieved an average correlation between human and machine scores as high as $r = 0.88$, although the performance of the system depended on the alignment accuracy.

In [9], quadratic polynomial equations were used to fit utterance-final pitch contours in utterances produced by two groups of Catalan-Spanish bilinguals. Slight differences in the contours in the two languages were observed between Catalan-dominant males and females using this methodology. In [10], a least square method using polynomials from the first to the seventh degree was used to describe the F0 contours for monosyllabic utterances of a tonal language, Yorùbá. The best approximation was achieved by the third degree polynomials, and the worst performance was with seventh degree polynomials.

In addition to F0 features based on polynomial fitting, simple descriptive statistics have also been shown to be a good indicator of prosodic proficiency. For example, [11] investigated the differences in pitch contours of yes/no questions between Russian native speakers and Finnish L2 speakers of Russian, and evaluated the correlation between native speaker perceptions of the non-native utterances and various pitch measurements, including maximum, mean, standard deviation, range , etc. This study found only weak correlations between these F0 measurements and native speaker ratings (the highest was $r = 0.36$ for the standard deviation of F0 measurements). Other studies have found that pitch range can be a useful feature for predicting accent ratings [12] and detecting cross-language differences [13]; this feature was also included in our feature set.

In the current study, we model F0 contours from a global perspective using contour features calculated based on a polynomial fit at the utterance level, as opposed to frame-by-frame values as was done by [8]. Following [10], we examine polynomial functions from three to seven degrees. We also combine these more complex contour-based metrics with simple descriptive statistics of F0 measurements.

3. Data

In order to investigate the prosodic characteristics of utterances produced by native and non-native speakers, we use a corpus that consists of 9,092 responses from non-native speakers from two different countries (China and India) and 4,218 responses from native speakers of US English. The distribution of the speakers by gender and country is shown in Table 1.

The non-native responses were drawn from a pilot administration of an international language assessment of English language proficiency for non-native speakers in India and China. The goal of the assessment is to evaluate the ability of non-native speakers to produce intelligible and fluent speech. For this study we selected responses to a Sentence Repeat task for which the test takers heard a short sentence as an audio stimulus, and were then asked to repeat the sentence verbatim in 10 seconds. The Sentence Repeat task was selected since the responses typically consist of a single intonational phrase. Because each stimulus passage generally consists of a single clause; the responses can thus be modeled relatively accurately with a single prosodic contour. The participants in China spoke Mandarin Chinese as their first language (L1), and the participants in India are represented by a diverse range of L1s.

The native speech corpus used in this study consists of spoken responses from speakers representing all major North American dialect regions; this corpus was collected independently of the non-native corpus described above. The speakers were asked to complete the same task as the non-native speakers. In total, 48 distinct Sentence Repeat prompts were included in the data set, and each speaker in the corpus responded to a total of 16 prompts.

Table 1: *Number of responses in each subset used in the nativeness classification experiment*

	Female	Male	Combined
Native	2,940	1,278	4,218 (31.7%)
China	2,691	2,051	4,742 (35.6%)
India	2,395	1,955	4,350 (32.7%)
Total	8,026 (60.3%)	5,284 (39.7%)	13,310

4. Methodology

4.1. Extraction of F0

For this study, we extracted F0 measurements for each response in the corpus using the auto-correlation F0 estimation algorithm in Praat [14]. Post-processing was done on the F0 measurements to correct implausible F0 jumps and interpolate over unvoiced regions with smoothing using a Butterworth filter with a normalized cut-off frequency equal to 0.1. The F0 measurements were extracted using a two-pass approach that optimizes the pitch floor and ceiling parameters in order to reduce pitch halving and doubling errors [15]. Finally, all raw F0 measurements were normalized to z-scores based on the means and standard deviations of all F0 measurements from all utterances from each speaker.

4.2. Features

The following 48 pitch contour features, which are all computed based on the normalized F0 measurements, are used in this study:

1. Basic statistical variables: min, max, mean, median, standard deviation, and range of the normalized pitch values, denoted as minF0, maxF0, meanF0, medianF0, stdevF0, rangeF0, respectively (6 features).
2. The coefficients of polynomial fitting functions from first degree to seventh degree, denoted as order-α-coeff-β, where $\alpha \in$ [1,7], $\beta \in$ [1,α+1]. The coefficients are numbered in descending order, e.g., order-3-coeff-1 stands for the coefficient of the highest degree of the cubic function (x^3). The best-fitting polynomial function of each degree was determined based on the standard deviation of residuals (35 features).
3. the standard deviation of residuals for each polynomial fitting function, denoted as std-α, where $\alpha \in [1, 7]$. e.g. std-2 refers to the standard deviation of residuals for the quadratic polynomial fitting function (7 features).

4.3. Experimental Design

We used the Random Forests algorithm [16] implemented in WEKA [17] to classify all responses into native and non-native using the 48 features described in the previous section. We evaluated the performance of the model for nativeness classification by using 10-fold cross-validation over the entire corpus. The number of trees in the forest was tuned from 10 to 100, with a step of 10 trees to determine the optimal value. The models are evaluated using weighted F_1 score. The weighted F_1 score was computed as follows: $F_{weighted} = \sum_{i=1}^{2} w_i * F_i$, where i indicates the class (native or non-native), w_i denotes the proportion of responses in class i and F_i indicates the F_1 score computed for this class. The majority baselines for each experiment were computed as weighted F_1 score for classification results based on the class label only.

We performed three groups of experiments. First, we performed classification using all 48 features described above with the full set of polynomial coefficients ranging from 1 to 7 degrees. We refer to this feature set as "Poly1-7". Since F0 contour patterns may differ between genders, we also explored this potential difference by conducting classification experiment separately for male and female speakers. Second, we reduced the feature set to a smaller one which only contains the polynomial functions of the first three degrees and basic statistics (18 features in total) and re-ran the same types of experiments. We refer to this second feature set as "Poly1-3". Third, we compared text-dependent models trained on native and non-native responses to the same item with text-independent models trained on multiple items.

5. Results

5.1. Text-independent classification

Table 2 presents the nativeness classification results in terms of weighted F_1 score achieved by the "Poly1-3" and "Poly1-7" feature sets over different subgroups: female-only, male-only and combined. The models were trained and evaluated on combined data from all 48 different prompts in our corpus.

Table 2: *Weighted F_1 score for the nativeness classification experiment using two different feature sets for gender-specific and combined groups*

Gender	Baseline	Poly1-7	Poly1-3
Female	0.492	**0.784**	0.764
Male	0.654	0.799	**0.800**
Combined	0.554	**0.796**	0.783

Table 2 shows that the classification performance does not decrease substantially when only polynomials of the first three degrees are used instead of seven. This result indicates that the the basic descriptive statistics and lower-degree polynomial parameters can capture most of the characteristics of the pitch contours. We also see larger relative improvement over the baseline for the female subgroup than the male subgroup for both sets of models.

5.2. Text-dependent classification

In order to investigate whether the F0 contour features would perform better in a text-dependent context, we subdivided the data into groups of responses to each of the 48 Sentence Repeat prompts in the data set. The average number of responses in each group was 277.

We next trained a text-dependent model for each of the 48 prompts using the "Poly1-3" feature set and generated predictions for each prompt separately. All experiments were performed on the combined data set since the size of the subset did not allow training gender-specific models. In order to expedite training on so many datasets, we fixed the tree number of Random Forest at 70, which was the most reliable number on average in other trials. We then pooled all these predictions together and computed the combined F -score over the whole data set. We next compared this F_1 score with the F_1 score from text-independent model presented in 5.1.

The results are summarized in Table 3. As shown in the table, the text-independent approach achieved better performance than the text-dependent approach when 10-fold cross validation was used. This is likely due to over-fitting in the text- dependent configuration since the data sets for the 48 individual prompts are much smaller than the pooled data set in the text-independent configuration and therefore the data sets generated during 10-fold cross validation were too small to train reliable models.

Table 3: *Weighted F_1 score for nativeness classification using text-dependent and text-independent models (Baseline: 0.554).*

Model	10-fold
Text-independent	0.783
Text-dependent	0.741

5.3. Attribute Analysis

In order to determine the most important features in the differentiation of native and non-native speech, we used the correlation-based feature subset selection (CFS) algorithm [18] as the feature evaluator, combined with the Best First search method [17] on both the "Poly1-7" and "Poly1-3" feature sets, under the text-independent condition. Table 4 presents the results of this feature selection approach. We find that the following features are most commonly selected as important across the different conditions: rangeF0, order-1-coeff-1, order-1-coeff-2, order-2-coeff-1, order-3-coeff-1. It is not surprising that rangeF0 plays a vital role in classification since it gives us a basic understanding of the amplitude of the pitch contour. The highest-degree coefficients distinguish between rising and falling contours and determine the overall shape of the pitch contour. In addition, minF0 was selected for both of the male groups but not the female groups.

Finally, we repeated the classification experiments based on the selected subsets of features. The results of nativeness classification using the selected features are presented in Table 5. This table shows that if we only use the subset of the most significant features, the model performance is lower than when using the full feature set, but still achieves weighted average F_1 scores higher than 0.7.

5.4. Gender and Country of Residence

The final set of experiments investigated whether the model performance depends on speakers' country of origin. We conducted separate classification experiments for non-native speakers from China and India using the same procedure as in Section 5.1. We split the non-native data into two parts based on

Groups	Poly1-7 Method	Poly1-3 Method
Combined	rangeF0 order-1-coeff-1 order-1-coeff-2 order-2-coeff-1 order-3-coeff-1 order-4-coeff-1 order-5-coeff-1 order-6-coeff-1 order-7coeff-1	n-rangeF0 order-1-coeff-1 order-1-coeff-2 order-2-coeff-1 order-3-coeff-1
Male	minF0 rangeF0 order-1-coeff-1 order-1-coeff-2 order-2-coeff-1 order-3-coeff-1 order-4-coeff-1 order-5-coeff-1 order-6-coeff-1 order-7-coeff-1	minF0 rangeF0 order-1-coeff-1 order-1-coeff-2 order-2-coeff-1 order-3-coeff-1
Female	rangeF0 order-1-coeff-1 order-1-coeff-2 order-3-coeff-1 order-4-coeff-1 order-5-coeff-1 order-6-coeff-1 order-7-coeff-1	rangeF0 order-1-coeff-1 order-1-coeff-2 order-2-coeff-1 order-3-coeff-1

Table 4: *Feature selection of Poly1-7 method and Poly1-3 method.*

Table 5: *Nativeness classification results using the subset of features based on feature selection*

Gender	Baseline	Poly1-7	Poly1-3
Female	0.492	**0.723**	0.709
Male	0.654	**0.759**	0.747
Combined	0.554	0.720	**0.742**

the speaker's country of origin, i.e. China or India. Then, we repeated the experiments described in Section 5.1 for the China group vs. the entire set of native data and the India group vs. the entire set of native data. The results are presented in Table 6 and Table 7, respectively. In agreement with the results reported in 5.1, the models showed larger improvement over the baseline for the female-only group than for the male-only group, for speakers from both China and India. For the male-only group and the combined data set, the models performed slightly better for Chinese speakers. In both experiments, the Poly1-7 method slightly outperformed Poly1-3, which is consistent with the results presented for the country-independent in Table 2.

Table 6: *Weighted F_1 score for the nativeness classification experiment using two different feature sets for gender-specific and combined groups (China vs. Native)*

Gender	Baseline	Poly1-7	Poly1-3
Female	0.358	**0.790**	0.778
Male	0.470	**0.808**	0.786
Combined	0.366	**0.802**	0.785

Table 7: *Weighted F_1 score for the nativeness classification experiment using two different feature sets for gender-specific and combined groups (India vs. Native)*

Gender	Baseline	Poly1-7	Poly1-3
Female	0.392	**0.780**	0.753
Male	0.456	**0.767**	0.751
Combined	0.342	**0.773**	0.761

6. Discussion

In this paper, we modeled raw pitch contours using linguistically-naïve models for the task of discriminating between native and non-native speakers of English. Using sentence-level pitch contour features, we employed Random Forests to conduct binary classification between the native and non-native sets under both text-dependent and text-independent settings. Contrary to our expectations, the experimental results showed that our models worked better in a text-independent context, which achieved a weighted F_1 score of 0.783 compared to a weighted F_1 score of 0.741 for the text-dependent models. This result, however, could potentially be due in part to the fact that the text-dependent models had substantially smaller training sets (since the data set was divided into 48 partitions for the text-dependent experiments); subsequent experiments with training sets of equal size between the text-dependent and text-independent configurations would be required to answer this question definitively. We also found that there is no substantial drop in the nativeness detection performance between models trained on polynomial functions on first seven degrees of polynomials and the models trained on only the first three degrees of polynomials. In other words, basic descriptive statistics and the lower-degree (i.e., Poly1-3) polynomial parameters can capture the primary characteristics of the pitch contours.

Furthermore, we highlighted the principle attributes for different feature sets and found that the highest-order polynomial coefficients of 1- to 7- degree fitting functions, constant polynomial of 1-degree fitting function, the range of normalized F0 are the best predictors of nativeness across all the three groups (female, male, combined).

These experimental results demonstrate the potential for exploiting differences in pitch contour features between native and non-native speakers. Since the speakers from India in this study represented a range of divergent L1 backgrounds, future work should investigate finer subdivisions among these speakers in order to more clearly demonstrate the effect of L1 on the approach. Our experiments suggest the general usefulness of the polynomial modeling of pitch contours in distinguishing non-native speech from natives, which could be applied as the pre-processing modules in computer-aided prosody assessment scenario, automatic pronunciation tutoring system, or prosody generation for speech-to-speech translator (such as in [19]). In the future, it will likely be beneficial to also extract the mean pitch values for each phoneme, and then extract the contour fitting parameters based on these values at the phoneme-level, as opposed to the frame-level results.

7. References

[1] S. Winters and M. G. O'Brien, "Perceived accentedness and intelligibility: The relative contributions of F0 and duration," *Speech Communication*, vol. 55, no. 3, pp. 486–507, 2013.

[2] X. He, J. Hanssen, V. J. van Heuven, and C. Gussenhoven, "Mandarin-accented fall, rise and fall-rise F0 contours in Dutch," in *Proceedings of Speech Prosody*, 2012.

[3] A. 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, pp. 600–603.

[4] C. G. Clopper and R. Smiljanic, "Effects of gender and regional dialect on prosodic patterns in American English," *Journal of Phonetics*, vol. 39, no. 2, pp. 237–245, 2011.

[5] J. Tao, K. Evanini, and X. Wang, "The influence of automatic speech recognition on the performance of an automated speech assessment system," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2014.

[6] J. Cheng, "Automatic assessment of prosody in high-stakes English tests." *Proceedings of Interspeech*, pp. 1589–1592, 2011.

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

[8] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Communication*, vol. 52, no. 3, pp. 254–267, 2010.

[9] M. Simonet, "Intonational convergence in language contact: Utterance-final F0 contours in Catalan-Spanish early bilinguals," *Journal of the International Phonetic Association*, vol. 41, no. 02, pp. 157–184, 2011.

[10] M. A. Fayemiwo and O. A. Odejobi, "Computational study of fundamental frequency of standard Yoruba monosyllabic utterances," *The International Journal of Computer Science and Communication Security*, 2014.

[11] R. Ullakonoja, "Pitch contours in Russian yes/no questions by Finns," in *Proceedings of Speech Prosody*, 2010.

[12] O. Kang, "Salient prosodic features on judgments of second language accent," in *Proceedings of Speech Prosody*, 2010.

[13] M. Urbani, "The pitch range of Italians and Americans. a comparative study." Ph.D. dissertation, University of Padova, 2003.

[14] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[15] K. Evanini, C. Lai, and K. Zechner, "The importance of optimal parameter setting for pitch extraction," in *Proceedings of Meetings on Acoustics*, vol. 11. Acoustical Society of America, 2011.

[16] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The Weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[18] M. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1998.

[19] P. D. Agüero, J. Adell, and A. Bonafonte, "Prosody generation in the speech-to-speech translation framework," in *Proceedings of Speech Prosody*, 2006.