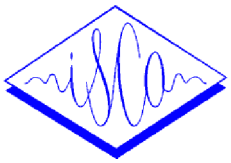
ISCA Archive
http://www.isca-speech.org/archive



Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

# TESTING SUPRASEGMENTAL ENGLISH THROUGH PARROTING

*Joseph Tepperman, Theban Stanley, Kadri Hacioglu, and Bryan Pellom*

Rosetta Stone Labs
Boulder, Colorado, USA
{jtepperman,tstanley,khacioglu,bpellom}@rosettastone.com

## ABSTRACT

Parroting exercises in a foreign language are designed to make a student's speech more native-like through imitation of specific native speech templates. In this paper we describe novel template-based methods for automatically estimating subjective scores for both intonation and rhythm in nonnative English. In terms of accuracy when automatically classifying a parroting speaker as a native or a learner, experimental results show that these new rhythm and intonation scores outperform similar baselines from nonnative speech assessment literature, and that they offer complementary discriminatory information when combined with automatic segment-level pronunciation scores, reaching a maximum classification accuracy of 89.8% on a corpus of parroting exercises. This suggests the general usefulness of these new scores in automatically assessing nonnative pronunciation in a computer-assisted pronunciation practice scenario.

***Index Terms***— nonnative speech, pronunciation evaluation, suprasegmental features, second-language acquisition

## 1. INTRODUCTION

Not unlike talking parrots, students of a second language (L2) can learn native-like pronunciation by imitating the speech they hear. *Parroting* is the attempt on the part of a language learner to reproduce any aurally-transmitted aspect of L2 pronunciation, in an effort to sound more native-like and eliminate the influence of their own native language (L1). This can encompass pronunciation on multiple linguistic levels - both segmental and suprasegmental imitation. It can also happen implicitly, as in complete immersion in a foreign country, or explicitly, as in a language practice scenario, in response to spoken or recorded prompts designed to elicit repetition.

The paradigm of student imitation of audio prompts is pervasive throughout second-language instruction [1, 2]. It has proven to be an effective way to teach phonological, suprasegmental, and conversational elements of an L2. Advanced learners can also benefit from parroting exercises, in learning to produce subtle contrasts between close phonemes, or acquiring the rhythm and intonation of a second language (a task usually reserved for students who have already mastered basic vocabulary and segment-level pronunciation). This paper will specifically address the problem of automatically scoring suprasegmental practice for Japanese learners of English.

To assess a student's accuracy in parroting an audio prompt, we propose a template-based scoring method. By "template-based" we mean a direct comparison between any input student utterance and the native speech audio prompt that elicited it (the reference "template") by way of some kind of distance metric, but without any model abstraction beyond that. This seemed attractive for evaluating parroting for several reasons. For one, intonation and rhythm are highly variable in English, even among native speakers [3]. The "tune" to which a phrase is set depends on the speaker's semantic intent, and also on their speaking style and emotional content. If a student is encouraged to imitate a specific prompt's suprasegmental realization, then that very prompt must serve as a reference. Though this limits the assessment to phrases that have been pre-recorded, the method is inexpensive computationally (it requires no training) and in terms of transcription, requiring no specialized prosodic annotation.

In this paper we will describe some novel template-based methods for estimating both intonation and rhythm scores. These proposed methods will be assessed alongside baselines from nonnative speech assessment literature, and all methods will be compared in terms of performance in automatically classifying a parroter as an English native or learner. It is our goal to show improved discrimination over the baselines and to demonstrate that these suprasegmental scores offer complementary information when combined with segment-level scores. The next section describes the corpus of parroting recordings used here. Following that is a description of the proposed methods and baselines. Finally, we will discuss the results of experiments in automatic speech classification.

## 2. SPEECH CORPUS

As mentioned in Section 1, this paper is focused on the task of evaluating suprasegmental parroting produced by Japanese students of English. The speech data used in all the exper-

[1]Thanks are due to Daniel Stutzman for implementing the data collection, and to all the Rosetta Stone employees who participated in it.

| | *Japanese Learners* | *Native English* | *Reference Prompts* |
|---|---|---|---|
| *Female Speakers* | 51 | 5 | 2 |
| *Male Speakers* | 44 | 11 | 2 |
| *Total Phrases* | 1478 | 1592 | 100 |
| *Total Hours* | 1.3 | 1.4 | 0.1 |

**Table 1**. Speech corpus statistics.

iments below is divided into three sub-corpora: Japanese learner parroting, Native English speaker parroting, and reference audio prompts. All recordings were made with 16-bit resolution, sampled at 16 kHz, and all speakers parroted some subset of 100 English phrases, hand-selected to represent a variety in length, subject matter, and suprasegmental content. Statistics about these three sets are given in Table 1.

The reference prompts were produced by 4 professional voicers (2 male, 2 female) and were previously used as actual prompts in the Rosetta Stone Version 3 American English product. These recordings were processed and edited (including multiband compression and equalization) to maximize intelligibility. In some sense these pronunciations are artificial - they are enunciated much more deliberately and dramatically than in ordinary speech or even formal reading. But as prompts they are appropriate, with intonation, rhythm, and phonetic quality unambiguous for purposes of imitation.

The learner parroting recordings were collected from native Japanese speakers who also live in Japan. Though none of them were fluent in English, the speakers encompassed a wide range of English proficiency. All recordings selected for these experiments were checked to ensure that they were at least devoid of noise and that they were in-grammar, i.e. that the student produced all of the words in the prompt. These learners were not explicitly told to parrot the reference prompts suprasegmentally. Even so, the influence of the prompts' intonation and rhythm were palpable, with many learners going as far as to take on the professional voicer's emphatic style.

To investigate the upper bound on parroting accuracy, we also collected recordings from native speakers of English who are employees of Rosetta Stone. Each speaker was explicitly told to try and match the voicer's rhythm and intonation, and they were allowed to listen to the prompt and record their own version as many times as they wanted. None of them were professional voicers. Though these conditions were not the same as in the Japanese collection, this yielded two very different data sets: one with presumably proficient native parroting, and one with learner parroting of diverse proficiency.

## 3. METHODS AND METRICS

To estimate scores for suprasegmental parroting, we begin by automatically segmenting the student's speech into syllables and generating segment-level acoustic scores. With this syllable segmentation we compare the student's segmentation to that of the reference prompt, and generate an appropriate score for rhythm similarity. Then we estimate the student's fundamental frequency (f0) contour, and compare it to that of the prompt - an intonation similarity measure is calculated from there. No energy-based scoring is done because the prompt audio was processed such that the energy was more or less uniform throughout, and so it can't serve as a good reference. These steps are explained in detail in this section.

### 3.1. Speech Segmentation

Automatic speech segmentation was done using forced Viterbi decoding of the target utterance using Rosetta Stone's proprietary speech recognition system. The segmentation process provided both word-level and phoneme-level alignments of the speech data. The decoded sequence of phonemes was then chunked into syllables based on each word's expected syllabification according to a pronunciation dictionary. The decoding grammar allowed for possible word deletion and silence insertion, to be expected in learner speech. Phonemes were assigned a pronunciation score based on a standard likelihood ratio. These scores were aggregated into an overall score for an utterance's segmental pronunciation.

### 3.2. Rhythm Measure

To compare the rhythm of the prompt and student utterances, the number of syllables for comparison needed to be identical for both. Due to pronunciation variants in the recognition lexicon, some student phrases may have been decoded with a different number of phonemes or even syllables from the template. For example, the word "temperature" (which occurred in our corpus) can be pronounced either as /tɛmpɚʧɚ/ with 3 syllables or as /tɛmpɚəʧɚ/ with 4. In cases where the number of syllables did not match, we backed off to word-level rhythmic analysis only for the word(s) with the differing number of syllables. Also, to compensate for differences in speaking rate, the template's syllable durations were linearly scaled so that the total duration would match that of the student.

The rhythm of speech is characterized not just by the durations of segments but by their contrast - between strong and weak, long and short, stressed and unstressed [3]. A common way of quantifying rhythmic contrast is the Pairwise Variability Index (*PVI*) [4, 5]; it is essentially the mean difference in duration between pairs of adjacent units (usually syllables) over an utterance. For directly comparing two speakers' rhythms, we propose a measure called the Pairwise Variability Error (*PVE*). Given a sequence of student durations $\{s_1, s_2, \ldots, s_N\}$ (most of which will represent syllables, but sometimes words if the number of syllables did not match the reference), and a sequence of corresponding reference durations $\{r_1, r_2, \ldots, r_N\}$, our rhythm score is defined as

$$PVE = \frac{\sum_{i=2}^{N} \sum_{m=1}^{min(M,i-1)} |(s_i - s_{i-m}) - (r_i - r_{i-m})|}{\sum_{i=2}^{N} \sum_{m=1}^{min(M,i-1)} |s_i - s_{i-m}| + |r_i - r_{i-m}|} \quad (1)$$

It sums up the "difference of differences" between pairs of syllables in the student and reference utterances, and then normalizes by the total absolute difference. If the student and reference durations are nearly equal, this score tends toward zero; as their difference approaches infinity, this score approaches 1. The $m$ variable is the rhythmic context - an integer $m \geq 1$ - which allows for comparisons between distant syllables (i.e. $M$ is the maximum distance, in syllables, considered for comparison). This is an enhancement over the *PVI* which only accounts for adjacent pairs. The idea is that the difference in duration between a pair of distant syllables might be more important, as a perceptual correlate of nativeness, than the difference between a pair of adjacent syllables.

### 3.3. f0 Contour Processing and Intonation Scoring

Fundamental frequency (f0) contours for all utterances were estimated using an autocorrelation-based maximum a posteriori method similar to the one presented in [6]. As with the rhythm score, some adjustments were necessary to ensure that the student and reference f0 contours were directly comparable. This was achieved by warping the reference contour to the student's length, phoneme-by-phoneme. Phoneme-level warping and concatenation ensured that the reference's voiced and unvoiced regions would roughly line up with those of the student. In the event that the number of phonemes did not match, warping backed off to the syllable level (but only for the syllables that did not match) and if the syllables were not the same number, then warping backed off to the word level. This warping was done using linear interpolation of the f0 contours - the template was warped to the student's durations and not vice-versa so as not to distort what the student produced. After warping, the similarity of the two contours was calculated as the correlation of the frames that were voiced for both the student and the reference. Correlation is a common measure for f0 contour similarity [7, 8], and was used here because it is insensitive to inter-speaker differences in f0 range; the unvoiced frames were ignored rather than interpolated so as not to put words in the student's mouth, so to speak.

### 3.4. Baselines

The standard non-template measures we used for our rhythm baselines are defined identically to those in [4, 5]. They all apply to student speech only, and are defined as follows: the standard deviations of vowel, consonant, and syllable durations ($\Delta$**V**, $\Delta$**C**, and $\Delta$**S**), as well as normalized versions of these (**varcoV**, **varcoC**, and **varcoS**); the rate of speaking (**ROS**) in phonemes per second; the vowel durations as a percentage of each phrase (**%V**); and the *PVI* measures for vowels and consonants (**PVI-V** and **PVI-C**). The *PVI* is the inspiration for the *PVE* measure in Eqn. 1, and is defined as

$$PVI = 100 \times \left( \frac{1}{N-1} \sum_{i=1}^{N-1} \left| \frac{s_i - s_{i+1}}{(s_i + s_{i+1})/2} \right| \right) \quad (2)$$

| *M* = | *1* | *2* | *3* | *4* |
|---|---|---|---|---|
| *Accuracy %* | 68.0 | 68.1 | 69.4 | 69.0 |

**Table 2**. Performance of the *PVE* over different $M$ in Eqn. 1.

where $\{s_1, \ldots, s_N\}$ is the sequence of segment durations (either vowels or consonants). Sources such as [4, 5] have shown significant differences in these metrics between native and nonnative English speakers, as well as between languages hypothesized to have fundamentally different rhythms.

Similarly, many studies have used Dynamic Time Warping (DTW) to compare f0 contours in template-based intonation scoring [9, 10]. DTW is a standard algorithm that aligns two time sequences that may differ in length. According to some defined cost function, this alignment returns the minimum cost of warping the two sequences together. Following [9], we constrained our DTW so that phoneme boundaries between the student and reference sequences are required to match - this also makes it comparable to the phoneme-level warping proposed in Section 3.3. Likewise, our cost function is the Euclidean distance between the student and reference f0 values - for normalization, they are divided by their sum and the mean number of frames in the two sequences.

## 4. EXPERIMENTAL RESULTS

To reiterate the goals of this study, we are interested in demonstrating that these new rhythm and intonation scoring methods outperform baselines from the literature in classifying native vs. learner parroting. We also intend to show that these scores can offer improvements in classification when combined with the segment-level confidence measures introduced in Section 3.1.

Combining the recordings from Japanese learners and native English speakers mentioned in Section 2, classification of all phrases as native or learner was performed using a leave-one-speaker-out crossvalidation procedure. For the individual scores described in Section 3, an optimal threshold for classification at the Equal Error Rate (EER) was found through a stepwise search through every fold's training instances. Combinations of scores were also investigated using multiple regression (as in [4]), which solved for the coefficients $\{a_0, a_1, \ldots, a_c\}$ in a set of equations of the form

$$y = a_0 + (a_1 \cdot x_1) + (a_2 \cdot x_2) + \ldots + (a_c \cdot x_c) \quad (3)$$

where each training phrase is represented by one equation, $\{x_0, x_1, \ldots, x_c\}$ are the scores to be linearly combined, and $y$ is set to 1 for native speech instances and 0 for learner speech instances. After this linear combination of the scores, the optimal EER threshold was found in the same way as for individual scores. The *PVE* (defined in Eqn. 1) was assessed for $M = \{1, 2, 3, 4\}$ - these results are reported in Table 2, and the best value of $M$ was used in all experiments with combinations of features (results reported in Table 3).

| | *rhythm* | *intonation* | *rhythm + intonation* | *rhythm + intonation + segmental* |
|---|---|---|---|---|
| *baselines* | 80.3 | 65.9 | 75.0 | 88.7 |
| *new measures* | 69.4 | 73.5 | 80.0 | 89.8 |
| *all* | 78.2 | 78.3 | 81.9 | 89.9 |

**Table 3**. Percent accuracy for native/learner classification. Classifying all as "native" achieved 51.9%; *segmental* alone achieved 87.2%. Native listeners should approach 100%.

## 5. DISCUSSION

According to Table 2, the value of $M$ with the highest classification accuracy was $M = 3$. This indicates that context beyond adjacent syllables ($M = 1$) is important in determining a score that can most accurately discriminate between native and learner rhythm parroting - results at $M = 3$ were significantly better than at $M = 1$ with $p \leq 0.03$, using McNemar's test. The fall-off in performance at $M = 4$ suggests that rhythmic contexts beyond 3 syllables are not useful here.

The results in Table 3 show that the proposed measures are, in general, improvements over the traditional baselines. The exception is the ensemble of rhythm baseline scores - ten in all, as listed in Section 3.4 - which together achieve 80.3% accuracy over the single new rhythm measure's 69.4%. However, the best of the baseline rhythm scores (the **ROS**) managed only 64.2% accuracy on its own, and the other baseline scores performed below chance levels when used individually. This suggests that the new *PVE* score is better than any of the baselines alone, and that most of the baselines are ill-suited for this task unless used in combination. However, the baseline rhythm scores have the advantage that they do not require a reference prompt for score calculation.

With the segment-level score described in Section 3.1, all suprasegmental measures - both baselines and novel metrics - offered complementary information for improved classification. The improvement in performance using the combined new measures over the baselines (the rightmost column of Table 3) was not significant with McNemar's test, indicating that the combined new measures are at least as powerful as the baselines, but maybe not moreso. It is interesting to note that the segment-level score alone achieved 87.2% accuracy compared to the new measures' combined 80.0%. This shows that these suprasegmental measures have substantial discriminatory power in comparison to their segmental counterparts, a result rarely seen in related work. Ordinarily, in studies like [11], suprasegmental scores perform at only a fraction of a segmental pronunciation score's accuracy.

## 6. CONCLUSION

For automatically classifying an English speaker as a native or learner, the proposed intonation scoring method outperformed a similar baseline; the new rhythm score outperformed each baseline individually, but not when the baselines were combined through multiple regression. In combination with segment-level scores, suprasegmentals offered an improvement over either one alone. We can conclude that the proposed measures are improvements over the baselines in estimating these binary pronunciation scores, and that they capture aspects of pronunciation quality not seen on the segment-level. This suggests they are useful for evaluating parroting. Future studies should show correlation with subjective listener scores, beyond this binary classification.

## 7. REFERENCES


[1] M.G. Busà, "New perspectives in teaching pronunciation," in *From Didactas to Ecolingua*, pp. 165–182. 2008.

[2] M.W. Tanner and M.M. Landon, "The effects of computer-assisted pronunciation readings on ESL learners' use of pausing, stress, intonation, and overall comprehensibility," *Language Learning & Technology*, vol. 13, no. 3, pp. 51–65, October 2009.

[3] P. Ladefoged, *A Course in Phonetics*, Thomson, Boston, 5th edition, 2006.

[4] T.-Y. Jang, "Automatic assessment of non-native prosody using rhythm metrics: Focusing on Korean speakers' English pronunciation," in *Proc. of the 2nd International Conference on East Asian Linguistics*, 2009.

[5] P.K. Mok and V. Dellwo, "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English," in *Proc. of Speech Prosody*, Campinas, Brazil, 2008.

[6] J. Droppo and A. Acero, "Maximum a posteriori pitch tracking," in *Proc. of ICSLP*, 1998.

[7] R.A.J. Clark and K.E. Dusterhoff, "Objective methods for evaluating synthetic intonation," in *Proc. of Eurospeech*, 1999.

[8] J. Mostow and M. Duong, "Automated assessment of oral reading prosody," in *Proc. of AIED*, 2009.

[9] D.T. Chappell and J.H.L. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proc. of ICASSP*, Seattle, 1998.

[10] 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, no. 4, pp. 83–90, 2008.

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