Connecting Rhythm and Prominence in Automatic ESL Pronunciation Scoring

Emily Nava[1]*, Joseph Tepperman*[2]*, Louis Goldstein*[1]*,*
Maria Luisa Zubizarreta[1]*, and Shrikanth Narayanan*[1,2]

[1] Department of Linguistics, University of Southern California, USA
[2] Signal Analysis and Interpretation Laboratory, University of Southern California, USA
http://sail.usc.edu/
{emily.nava@,tepperma@,louisgol@,zubizarr@,shri@sipi.}usc.edu

Abstract

Past studies have shown that a native Spanish speaker's use of phrasal prominence is a good indicator of her level of English prosody acquisition. Because of the cross-linguistic differences in the organization of phrasal prominence and durational contrasts, we hypothesize that those speakers with English-like prominence in their L2 speech are also expected to have acquired English-like rhythm. Statistics from a corpus of native and nonnative English confirm that speakers with an English-like phrasal prominence are also the ones who use English-like rhythm. Additionally, two methods of automatic score generation based on vowel duration times demonstrate a correlation of at least 0.6 between these automatic scores and subjective scores for phrasal prominence. These findings suggest that simple vowel duration measures obtained from standard automatic speech recognition methods can be salient cues for estimating subjective scores of prosodic acquisition, and of pronunciation in general.

Index Terms: nonnative speech, prosody, rhythm, pronunciation scoring

1. Introduction

The perception of speech as native or non-native is the result of a complex of production-based factors. The present study examines rhythmic cues as characteristic of English prosodic acquisition, and investigates their inclusion as part of the automatic pronunciation scoring process.

The data presented here were collected as part of a prior study [1, 2] that addressed the question of prosodic proficiency in the speech of second language (L2) learners of English whose first language (L1) is Spanish. A control group of 23 native English speakers and a test group of 46 L2English/L1Spanish speakers participated in a dialogue task which was designed to elicit prosodic production patterns associated with a given information structure. Of particular relevance to the current discussion are those contexts with intransitive verbs where all the information was new, such as the question and answer pair:

- Q: What happened?
- A: A glass broke.

For this context, native English speakers consistently produced main phrase-level prominence (also known as the nuclear pitch accent) on the subject, 'glass.' While some L2English speakers also produced a native-like prosodic pattern, others did not, instead placing main prominence on the verb (i.e. 'A glass broke'). This was taken as evidence of prosodic transfer from Spanish, since in Spanish for the same context, main prominence would also fall on the subject, but the latter would be sentence-final, as afforded by flexible word order in Spanish:

- Q: ¿Qué pasó?
- A: Se rompió un vaso.

In fact, main prominence is always sentence-final in Spanish for contexts such as this, where the information being communicated in new, or not previously mentioned [3, 4, 5]. Therefore, a Spanish speaker's use of sentence-final phrasal prominence in English was found to reliably indicate a lack of English prosodic acquisition.

English and Spanish differ not only with regards to prominence placement at the phrasal level, but also where organization at the rhythmic level is concerned. Empirical studies in the area of rhythmic classification have resulted in the categorization of languages into rhythmic classes based primarily on durational measurements [6, 7, 8, 9], and not in the terms of isochrony for which the names "stress-timed" and "syllable-timed" were originally proposed [10, 11]. However, in this study we will retain the use of these names for the sake of expository convenience.

English and Spanish are ideal languages to compare in this respect, as English is considered "stress-timed" due to the presence of vowel reduction, varied syllable structure inventory including complex onsets and codas, and vowels in stressed syllables that are regularly longer than in unstressed syllables. Spanish, on the other hand, is considered "syllable-timed" and does not have vowel reduction, has a reduced syllable inventory in comparison with stress-timed languages, and the difference between stressed and unstressed vowels is not as great.

The question of vowel reduction is at the crux of the relationship between events at the rhythmic level and the phrasal level. In English, the foot is the operative rhythmic unit, where strong and weak syllables self-organize in an alternating pattern. This pattern is granted by the significant reduction of the vowel in the 'weak' syllable as compared to the vowel in the 'strong' syllable. However, in Spanish the difference between syllables is not as great, prompting some to postulate that the foot does not function as a rhythmic unit in Spanish [13, 14, 15]. The greatest durational contrast among vowels in Spanish is of the stressed syllable in phrase-final position - which, as previously mentioned, is also where main phrasal prominence falls in Spanish. We hypothesize that those speakers who have acquired phrasal prominence in English will have also acquired the English foot as a rhythmic unit. This implies the prediction that a contrast in strong and weak vowel durations can be observed between those speakers who have acquired phrasal prominence

Table 1: Amount of native and nonnative English data used in this study. P = content word primary stress, S = content word secondary stress, F = function word.

		vowels		
population	*speakers*	P	S	F
ENC	23	161	69	253
L2E+PP	10	63	27	99
L2E-PP	36	259	111	407

in English and those who have not.

The current study not only provides evidence that English and Spanish differ in these rhythmic regards, but it also grants insight as to what aspects of second language speech contribute to the judgment of its prosodic nativeness. We intend to demonstrate correlation between native-like English phrasal prominence and native-like English rhythm as measured through the relative durations of primary and secondary-stressed vowels. This has implications for automatic pronunciation scoring in language learning applications, since phrasal prominence, though a good objective indicator of a speaker's acquisition of native English prosody, cannot be estimated automatically very easily. Vowel durations, however, can be estimated with standard alignment of target acoustic models.

The full range of data used in this study will be discussed more definitively in [16].

2. Corpus and Transcription

From the dialogue task described in Section 1, each speaker's degree of English prosody nativeness was defined as the number of times (out of 8 stimuli) that they put the phrasal prominence on each sentence's subject. The 23 native English speakers (population *ENC*) scored between 6 and 8, with an average of 7.6; 10 of the L2English speakers scored 5 or above, and the remaining 36 fell below that, averaging a 1.3 score among them. From these scores the L2English speakers were split into two populations: those who had acquired English-like use of phrasal prominence (population *L2E+PP*) and those who had not (population *L2E-PP*). In addition to the dialogue task, participants were also recorded reading the phonetically-balanced passage, "The North Wind and the Sun". All L2English speakers also read the passage in Spanish, and data from a monolingual control group of 20 Spanish speakers were also included. Statistics about the size of this corpus are summarized in Table 1.

Automatic transcripts were generated automatically by aligning phoneme-level Hidden Markov Models (HMMs) using an iterative bootstrap training procedure similar to that outlined in [17]. These HMMs were trained on 39-dimensional MFCC feature vectors, with 3 hidden states and 16 Gaussian mixtures per state. The window length was a standard 25 msec and the frame rate was shorter than usual (5 msec) so as to make segmentation times as accurate as possible.

We began with prior knowledge of the target phoneme sequence for each audio file, but without segment-level boundaries. Initial models were trained using the Baum-Welch embedded re-estimation algorithm, ignoring phoneme segmentation times. With these preliminary models we decoded each target phoneme sequence in the data set, allowing for optional pauses at expected phrase boundaries. The resulting phoneme

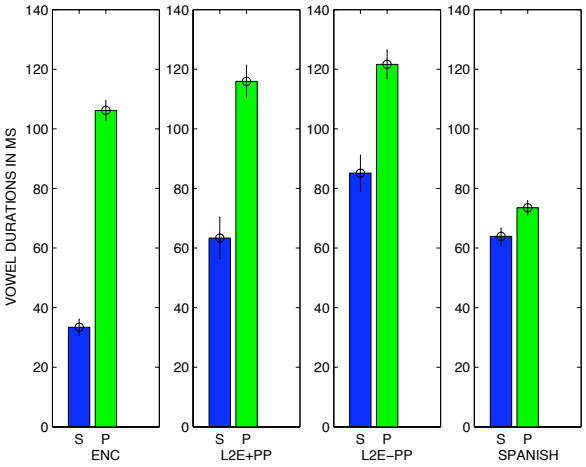

Figure 1: Normalized vowel durations for four speaker populations, in two contexts: content word primary stress (P), and content word secondary stress (S).

segmentation times were used to train new HMMs from scratch, this time using the hypothesized segmentation for model initialization with Viterbi alignment and then embedded re-estimation on each isolated phoneme (rather than over the whole sequence). Then the target sequences were decoded once more, and the new segmentation times were used to again train new acoustic models. This process of decoding and retraining was repeated for 5 iterations, at which point the automatic segmentations were found to be accurate with respect to a reference expert human segmentation.

We trained a separate set of HMMs for each of the three speaker populations. Over all ENC speakers and monolingual Spanish speakers there were 10.4 and 10.7 total minutes of speech available, respectively, and for all L2E speakers (both +PP and -PP) there were 53.8 minutes of speech available, since each speaker read both Spanish and English passages. The fully-trained monolingual English and monolingual Spanish models served to initialize those of the L2English speakers, for whom the pronunciation was highly variable and potentially drawing on both phoneme sets. For these reasons, decoding the L2English recordings also allowed for expected English pronunciations reflecting the influence of Spanish phonology. The recognition pronunciation lexicon included variants derived from Spanish letter-to-sound rules including, for example, the substitution of Spanish dental stops for English alveolar stops, or the possible lack of English-like aspiration in syllable-initial voiceless stops.

After automatic alignment, all durations extracted from segments of interest were normalized for speaking rate. These automatic segmentations were potentially inaccurate if the speaker paused at an unexpected place while reading the stimuli. In those cases, the alignment would include the pause as part of an abnormally long segmentation for the preceding phoneme. To eliminate these outliers, any voiceless sequence over 250 msec was considered a pause and subsequently removed from the analysis.

3. Statistics

Following [9], the data from the reading passage were used to calculate a voicing ratio [12] for each speaker. As expected,

English monolinguals as a group had a high voiceless-to-voiced ratio value, due to the presence of vowel reduction and complex consonant clusters, which can result in higher incidences of voiceless sequences in the speech stream. By contrast, values from the voicing ratio of the monolingual Spanish speakers were low. Statistical tests revealed that those speakers whose L2English was target-like for phrasal prominence also had voicing ratio values comparable to (not significantly different from) that of English native speakers ($p < .05$).

While these results are indeed encouraging as to the relationship between prosodic events at the phrasal level and those at the rhythmic level, the voicing ratio obscures the details of the nature of the value in the case of the L2 learners. It is with this in mind that we decided to look at vowel durations across all words in order to test the hypothesis laid out in Section 1 that those L2English speakers with native-like phrasal prominence would also have reduced vowels in secondary stress (i.e. "weak" syllable) positions, whereas those L2 speakers with Spanish prosodic transfer may or may not have reduced secondary-stress vowels.

A two-way ANOVA revealed a significant difference in durations ($p < .05$) in English between vowels belonging to different lexical classes (content words versus function words) and between vowels in content words with primary stress and those with secondary stress. However, in Spanish no significant difference was found among the vowels of any of the aforementioned word classes. This finding grants further support to the understanding that the hallmark of English rhythm lies in the difference between stressed and unstressed vowels (strong and weak elements, respectively, that make up the "foot" in English). However, rhythm in Spanish is not marked by such contrasts at the vocalic level.

Thus native speakers of Spanish who learn English as a second language are faced with the challenge of acquiring vowel reduction as they advance towards native-likeness in their L2 speech. Analysis of the vocalic production of L2English speakers revealed that those participants with native-like phrasal prominence (L2E+PP) in their English had the same patterns of significance in their durational differences as the native English speakers, while those without native-like prominence (L2E-PP) did not show these same significance patterns among vowel categories. See Figure 1 for a bar graph of mean vowel durations (normalized for speaking rate) in the four speaker populations: the significant difference between primary and secondary stress durations is visible for those populations with English phrasal prominence (ENC and L2E+PP), but not for those without (L2E-PP and SPANISH).

4. Automatic Pronunciation Scoring

The theory presented in Section 1 connected English prosodic acquisition manifested through phrasal prominence to English syllable rhythm manifested through vowel duration contrasts. This begged the question, can objective and automatically-measured vowel durations be used to make subjective judgments of prosodic acquisition? To begin to answer this, a correlation must be shown between these vowel durations and the prosodic scores explained in Section 2 that are supposed to be so indicative of English prosodic acquisition.

What is the best way to use these durations to characterize a contrast between "syllable-timed" or "stress-timed" production? The latter designation assumes that there is greater variation in vowel durations among lexical categories, while the former assumes that all vowel lengths are less variable. Figure 1 illustrates the mean vowel durations for each speaker population in the contexts of primary and secondary stress (P and S, respectively). Judging from these graphs, the defining characteristic of native English rhythm seems to be the significant difference in duration between vowel categories - this is not seen in the monolingual Spanish speech, with its insignificant variation in vowel durations.

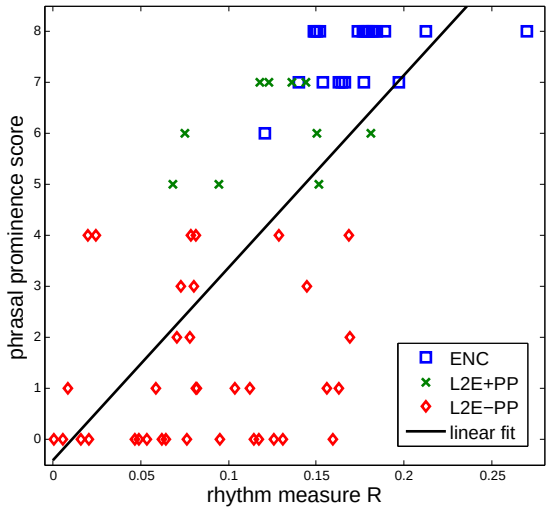

Figure 2: Scatterplot of speaker-level rhythmic and prosodic scores, split by population.

In keeping with this characterization is the most common measure for speech rhythm, the Pairwise Variability Index (PVI) [18]. It captures the mean of the normalized differences in duration between all pairs of adjacent segments (in this case the segments are vowel durations in adjacent syllables). Since we are working only with mean durations over a speaker, we define our PVI-inspired rhythm measure R as:

$$R = \frac{|P_n - S_n|}{(P_n + S_n)/2} \quad (1)$$

where, for speaker n, P_n is their mean primary stress vowel length, and S_n is their mean secondary stress vowel length.

With this rhythm measure, we achieve a correlation coefficient of 0.683 between R and the phrasal prominence scores. This is empirical evidence for a strong connection between English syllable rhythm and phrasal prominence. Figure 2 shows the scatterplot of rhythmic measure R versus phrasal prominence scores for the three speaker populations who read passages in English. Along the x-axis there is a visible separation between the L2E-PP and the ENC populations, indicating the usefulness of R as a measure of English nativeness. Since we are working with overall mean vowel durations across readings of "The North Wind and the Sun," the number of data points is equal to the number of speakers. Their context-dependent durations were normalized using the method described in Section 3.

In addition to the PVI-inspired rhythm measure above, we also tried using a linear combination of the mean primary and secondary stress durations per speaker (P_n and S_n) to predict the phrasal prominence score. With a leave-one-speaker-out crossvalidation training procedure, a trained linear regression on these durations resulted in a 0.604 correlation coefficient between phrasal prominence scores and the automatic scores predicted by the equation, again validating the theory connecting

rhythm and prominence. This trained regression equation was

$$PP_n = 10.5 \cdot P_n - 79.9 \cdot S_n + 8.7 \quad (2)$$

where PP_n was the phrasal prominence score for speaker n. This indicates that the secondary stress duration, S_n, is the more important of the two cues, though it is inversely proportional to PP_n. The implication is that, in nonnative English produced by Spanish learners, the secondary stress durations change more than primary stress durations when a speaker acquires English rhythm and phrasal prominence, and the mean durations illustrated in Figure 1 seem to support this.

5. General Discussion

The statistics of the corpus and the experiments in automatic scoring agree: syllable-level rhythm is connected to phrase-level prosody. Obtaining vowel durations from forced alignment of speech is relatively straightforward to do automatically. In a computer-aided language learning scenario, we would expect to have prior knowledge of the target phonemes to be aligned. If these durations can then be used to predict a more elusive phrasal prominence measure (as Section 4 demonstrates), then this method can be extended to estimating subjective scores for a speaker's overall level of prosodic nativeness, as well as their overall pronunciation quality on levels beyond just prosody. Furthermore, knowledge of which aspects of second language speech are more perceptually salient to native speakers allows for curriculum development design that can more precisely address the production stages of second language acquisition. The complexity of the cluster of aspects responsible for a judgment of speech as non-native can be teased apart, isolating those characteristics that represent the greatest contrast between the native and second languages. In the study presented here, the data clearly point to how a difference in vowel duration within word classes is not only a determinant factor for native-likeness, but more importantly that a connection exists between vowel duration at the rhythmic level and the production of prominence at the phrasal level.

6. Conclusion

The data presented here have important implications for the area of language development in two respects: a connection between rhythm and phrasal prominence has been established, and the relevance of vowel duration for the judgment of speech as native-like was likewise demonstrated. In future work in this area, perceptual tests will be needed to demonstrate a high correlation between subjective pronunciation quality and the measures of rhythm and prominence that are described in this study. A higher correlation between automatic scores and these subjective scores can be expected from incorporating additional rhythmic features and further durational cues such as voice onset time (VOT) or voicing ratio into a more sophisticated machine learning framework. Knowledge drawn from the above-described forced alignment technique can yield precise measurements of how native and non-native speakers differ regarding VOT for vowels. This in turn provides initial insight to the gestural components of rhythmic production even before examining gestural behavior.

7. References

[1] E. Nava and M.L. Zubizarreta, "Prosodic Transfer in L2 Speech: Evidence from Phrasal Prominence and Rhythm," *Proceedings of Speech Prosody 2008*, ed. P. Barbosa, S. Madureira, and C. Reis. Campinas, Brazil, 2008.

[2] E. Nava and M.L. Zubizarreta, "Deconstructing the Nuclear Stress Algorithm: Evidence from Second Language Speech." *The Sound patterns of Syntax*, eds. N. Erteschik-Shir and L. Rochman. Oxford University Press. In press.

[3] M.L. Zubizarreta, *Prosody, focus, and word order*. Cambridge, Mass.: MIT Press, 1998.

[4] J.M. Sosa, *La Entonación del Español*. Madrid: Ediciones Cátedra, 1999.

[5] J.A. Hualde, "Stress removal and stress addition in Spanish," *Journal of Portuguese Linguistics* 6:58-89, 2007.

[6] R. Dauer, "Stress Timing and Syllable Timing Reanalyzed," *Journal of Phonetics* 11:51-62, 1983.

[7] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition* 73:265-292, 1999.

[8] E. Grabe, and E. Low, "Durational Variability in Speech and the Rhythm Class Hypothesis," in C. Gussenhoven and N. Warner (Eds.), *Papers in Laboratory Phonology 7*. Berlin: Mouton de Gruyter, 2002.

[9] V. Dellwo, A. Fourcin, and E. Abberton, "Rhythmical classification based on voice parameters," *International Conference of Phonetic Sciences (ICPhS)*, 2007.

[10] K. Pike, *The Intonation of American English*. Second edition. Ann Arbor: University of Michigan Press, 1946.

[11] D. Abercrombie, *Elements of general phonetics*. Chicago: Aldine, 1967.

[12] E. Nava, L. Goldstein, E. Saltzman, H. Nam, M. and Proctor, "Modeling prosodic rhythm: Evidence from second language speech," *Acoustical Society of America annual meeting*, Miami, FL, November, 2008.

[13] F. Cummins and R.F. Port, "Rhythmic constraints on stress timing in English," *Journal of Phonetics*, 26(2): 145-171, 1998.

[14] I.M. Roca, "Secondary stress and metrical rhythm," *Phonology Yearbook*, 3:341-370, 1986.

[15] I.M. Roca, "Theoretical implications of Spanish word stress," *Linguistic Inquiry* 19, 393-423, 1988.

[16] E. Nava and M.L. Zubizarreta, "On the Nuclear Stress - Rhythm Connection: Evidence from Second Language Speech," forthcoming.

[17] S. Young et al. *The HTK Book.* [Online]. Available: http://htk.eng.cam.ac.uk/, 2002.

[18] P. Ladefoged, *A Course in Phonetics*. 5th Edition. Boston: Thomson, 2006.