Effects of Intonational Phrase Boundaries on Pitch-Accented Syllables in American English

Yen-Liang Shue[1], *Stefanie Shattuck-Hufnagel*[2], *Markus Iseli*[1],
Sun-Ah Jun[3], *Nanette Veilleux*[4], *Abeer Alwan*[1]

[1] Department of Electrical Engineering, University of California, Los Angeles
[2] Research Laboratory of Electronics, Massachusetts Institute of Technology
[3] Department of Linguistics, University of California, Los Angeles
[4] Department of Computer Science, Simmons College

yshue@ee.ucla.edu, stef@speech.mit.edu, iseli@ee.ucla.edu,
jun@humnet.ucla.edu, nanette.veilleux@simmons.edu, alwan@ee.ucla.edu

Abstract

Recent studies of the acoustic correlates of various prosodic elements in American English, such as prominence (in the form of phrase-level pitch accents and word-level lexical stress) and boundaries (in the form of boundary-marking tones), have begun to clarify the nature of the acoustic cues to different types and levels of these prosodic markers. This study focuses on the importance of controlling for context in such investigations, illustrating the effects of adjacent context by examining the cues to H* and L* pitch accent in early and late position in the Intonational Phrase, and how these cues vary when the accented syllable is followed immediately by boundary tones. Results show that F_0 peaks for H* accents occur significantly earlier in words that also carry boundary tones, and that energy patterns are also affected; some effects on voice quality measures were also noted. Such findings highlight the caveat that the context of a particular prosodic target may significantly influence its acoustic correlates.

Index Terms: prosody, voice quality, pitch accents, tonal crowding

1. Introduction

A challenging issue in spoken prosody is the difficulty of specifying the acoustic correlates of the phonological entities proposed in prosodic theories. It has not proven easy, for example, to determine the defining characteristics of various types of prominence, such as the phrase-level pitch accents and phrase-boundary-related tonal targets proposed in [1] and [2]. One difficulty is that these elements are often defined in relational terms; for example, the F_0 target associated with an H* accent for one speaker might correspond to the F_0 target of an L* accent for another. Another is that the relation between F_0 and a prosodic category can be complex; for example, the F_0 of a syllable that bears a downstepped H* accent will be lower than that of the preceding t-toned syllables, even though it is an H*. Other complexities include the fact that the F_0 peak of an H* may not occur within the accented syllable [3]; portions of the F_0 contour of an utterance may be distorted or obscured by segmental effects (such as stiffening of the vocal folds to prevent voicing, affecting their rate of vibration); and regions of irregular pitch may interfere with estimation of the vibration rate.

In addition to these challenges, the context in which a given prosodic element occurs can have significant effects on its acoustic realization. This concept is familiar from segmental phonology, where the effects of both position in structure (e.g. word- or syllable-initial vs. final) and adjacent segments (e.g. non-aspirated voiceless stops after tautosyllabic /s/) have long been observed. The picture is complicated for prosody, however, because several prosodic characteristics can be realized on the same word or even on the same syllable. For example, a lexically-stressed syllable may carry a pitch accent or not, and a pitch-accented word may also carry a boundary tone marking the end of an intonational phrase. Thus in order to determine the basic acoustic correlates of a prosodic element, as well as to determine how those correlates vary in the presence of other prosodic elements, it is important to control and systematically manipulate the prosodic context in which the target element occurs. In an earlier study [3] we compared the acoustic correlates of lexical stress when it was accompanied by H* or L* pitch accents or no pitch accent. In the present study we carried out a conceptually similar manipulation to determine the effects of concurrent boundary tones on H* and L* pitch accents. Our assumption was that in order to measure the acoustic correlates of a given type of intonational category accurately, it is necessary to remove from the context both concurrent and adjacent intonational targets that might significantly affect its realization. This approach lays the groundwork for revealing the systematic effects of those contextual elements in later studies, and for testing the predictions of various competing prosodic theories.

In [3] we noted a number of differences between pre-nuclear and nuclear H* and L* accents. However, in that study it was not possible to be certain that the differences were explicitly due to the intonational-phrase-medial vs. final position of the accents because of two confounds: the two types of accents were produced on two different lexical items (which had different stress patterns and contained different vowels), and the nuclear-accented word was the last word in the utterance, so that it carried boundary-related tones as well the accent. Thus it might have been these factors, rather than the pre-nuclear/nuclear positional contrast, that were responsible for the differences found. In the present study, we controlled for these factors, keeping the vowels in the accented syllables constant and systematically varying the presence vs. absence of the boundary tone, and early vs. late position in the intonational phrase, to examine the effects of boundary-tone context and position on the realization of H* and L* pitch accents. We examined four potential acoustic correlates of the accents: F_0 contour, energy, duration, and the voice quality measure H_1^*-H_2^* (related to open quotient [5]).

2. Data

For this study we elicited a corpus of spoken utterances with specified pitch accent and boundary tone locations and types. The elicitation stimuli consisted of the 2 sentences *Dagada gave Anne a dada* and *A dada gave Anne dagadas*, with a single pitch accent (H* or L*) produced either on the early target word or the late one, in either declarative or interrogative sentences. The nonsense words *dada* and *dagada* were used to ensure that the lexically-stressed syllables carrying the pitch accents (*da-* and *–ga-,* respectively) had the same vowel in all cases, avoiding any vowel-specific effects. The declarative and interrogative forms of the sentences were included to elicit use of the phrase-final tone sequences L-L% and H-H%, respectively. The same eight sentences were also recorded with the word *daily* added at the end of the sentence, to carry the boundary tone; this allowed us to determine the effects on the pitch accent realization in the late target word when the boundary tone was moved to a following word. Each of these 16 sentences was elicited with a prompt question or statement, to ensure the correct placement of the tones. For example, to elicit a H* tone on the early target word *dagada,* and a L-L% boundary tone on the unaccented late target word *dada*:

Prompt: *Was it Dagada or Daguda that gave Anne a dada?*

Response: *Dagada gave Anne a dada.*

Recordings were made for 10 native speakers of American English (5 male/5 female) between the ages of 17 and 30. For each sentence, 5 repetitions were recorded, for a total of 800 utterances. The signals were recorded in a sound booth at an effective sampling rate of 16 kHz. Manual segmentation of the target words *dada* and *dagada* from their context and segmentation of their main stress vowels were performed.

3. Methods

Analysis of the target words and main-stress syllables of the elicited utterances included estimation of F_0, energy contours, duration and H_1^*-H_2^* (related to voice quality [5]).

F_0 was estimated using the STRAIGHT algorithm [4], and polynomial fitting [3] was performed on the F_0 values over the target words to smooth the contours. Each utterance was manually checked to ensure that the polynomial fitting accurately represented the raw values. The maximum and minimum values were calculated from the smoothed contours and normalized to each speaker's mean value, calculated from all of the speaker's utterances. The duration of each main-stressed vowel was obtained from manual segmentation. Energy measures were calculated using a dynamic window size to take account of the effects of F_0. That is, for a particular point in time, the window size is set to 3 pitch periods as determined by F_0. The measure is then normalized to the mean energy of the particular utterance.

The measure H_1^*-H_2^*, related to open quotient [5] which influences perceived voice quality, is the difference between the first and second spectral harmonic magnitudes, corrected for effects of the first two formant frequencies [6]. The harmonic magnitudes were estimated from the speech magnitude spectra using F_0 information from the STRAIGHT algorithm. Formant corrections used the first two formant frequencies and bandwidths, estimated from the Snack Sound Toolkit [7] with the following settings: pre-emphasis factor of 0.96, window length of 25 ms, and window shift of 1 ms. The raw H_1^*-H_2^* values were then smoothed using Legendre polynomials of degree three to capture the general trends of the contours. Mean H_1^*-H_2^* values were calculated from the smoothed contours.

For all four measures, statistical analysis used two-way analysis of variance (ANOVA) tests using the software package SPSS (v13.0) with the two fixed factors speaker and tone (H/L), or speaker and boundary (yes/no, and if yes, H/L).

4. Results

For statistical analysis we focused on the vowels of the main-stress syllables of the target words, which had relatively clear boundaries. We distinguish between several properties of analyzed syllables: early vs late position of the target word in the utterance; for late position, boundary (bnd) vs. non-boundary (no-bnd) position; position of the lexically-stressed syllable in the target word (medial in *dagada*, initial in *dada*); accentedness (acc vs. no-acc) and if accented, whether the accent was H* or L*. We did not compare the effects of vowel type because we used the vowel /a/ in all target syllables.

Four types of positions are examined, illustrated here for declarative sentences with the target word *da**ga**da*, where the ***–ga-*** syllable is always stressed and can be accented or not:

1) no-bnd-early: *Da**ga**da gave Anne a dada.*
2) bnd: *A dada gave Anne da**ga**das.*
3) no-bnd-early-daily: *Da**ga**da gave Anne a dada daily.*
4) no-bnd-late-daily: *A dada gave Anne da**ga**das daily.*

Approximately the same number of tokens, 24-26, is included in each of the analysis cells. In all ANOVA tables of the results section, a significance level of $p \leq 0.001$ is regarded as statistically significant, and the partial η^2 (measure of effect size) is given in parentheses. Statistically non-significant results are marked with "No". Results for the words *dagada* and *dada* are separated, though in most cases they are similar.

4.1. F_0

Previous work [3] examined the F_0 contour in sentences like *Dagada gave Bobby doodads*, where the target word *dagada* always appeared early in the phrase and the target word *doodads* was always the final word of the utterance, so that it carried both the phrase accent and the boundary tone of the intonational phrase. Thus it was difficult to interpret the finding in that study that the F_0 extremum of the H*-related contour was often delayed beyond the end of the accent-related vowel in *dagada*, but not in *doodads*. (In contrast, the F_0 extremum of the L* contour was usually located in the middle of the main-stress syllable for both target words.) This result could have arisen from the differences in syllable structure of the two words, from differences in their main-lexical-stress vowels, from differences in their serial order in the utterance, or from differences in their location with respect to the phrase boundary. We hypothesized that the latter factor was critical. That is, the F_0 peak for an H* accent is consistently located in or just after the accented vowel, unless the need to realize other tonal targets on the same word causes the speaker to realize the peak earlier in the accented syllable, to make room for realization of the additional targets. To test this hypothesis, we analyzed the effect of serial position in the utterance (early vs. late) and of the presence/absence of a boundary tone on the same word (*bnd* vs. *no-bnd*) for the late position. We predicted that a) similar degrees of F_0 peak delay would be seen for both target words in both early and late positions, as long as a phrase boundary did not occur in the same word, and that b) lesser delay (earlier F_0 peaks) would be seen for both target words in late position if they were followed by an additional word *daily*, to

carry the boundary tone. Figure 1 shows F0 peak position for H* for a typical speaker producing *dagada* in three different contexts. Note that the *no-bnd-early(-daily)* and *no-bnd-late-daily* cases show similar temporal positions with respect to the accented vowel, i.e. the peaks are at 70-95% of the accented vowel. In contrast, the *bnd* cases show a systematically earlier peak location, at 50-60% of the vowel duration.

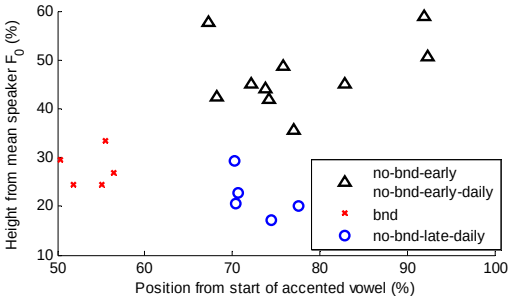

Figure 1: *Scatter plot for the target word* ***dagada*** *showing relative F_0 peaks for H* and their relative positions in the accented target vowel for a female speaker in three different contexts: 1. no-bnd-early/no-bnd-early-daily (triangles); 2. bnd (crosses); 3. no-bnd-late-daily (circles).*

Table 1 shows the F_0 results for the target word *dagada* for all subjects and all pitch accent/boundary tone types. The values in the *Position* row represent the mean F_0 peak/valley positions relative to the target vowel duration; 50% corresponds to the middle of the target vowel. The values in the *Height* row denote the mean F_0 peak/valley height relative to the speaker mean F_0 in percent; 0% corresponds to the speaker mean F_0 calculated over all the speaker's utterances.

Table 1: *Relative F_0 mean, standard deviation (std.) in parentheses, of H*/L* peaks/valleys and their positions for the target word* ***dagada*** *for male and female speakers. The statistical significance (s.s.) column shows the ANOVA results for the factor boundary condition (no-bnd-late-daily vs. bnd), while the s.s. row shows results for pitch-accent type (H* vs. L*).*

ΔF_0 mean (std) in %		Males			Females		
		no-bnd-late-daily	bnd	s.s. (η^2)	no-bnd-late-daily	bnd	s.s. (η^2)
H*	Position	78 (23)	62 (15)	Yes (.319)	86 (15)	67 (12)	Yes (.745)
	Height	10 (22)	3 (9)	Yes (.288)	26 (40)	24 (38)	No
L*	Position	53 (10)	49 (13)	No	57 (8)	60 (12)	No
	Height	-23 (12)	-23 (11)	No	-36 (10)	-34 (6)	No
s.s. (η^2)	Position	Yes (.564)	Yes (.442)		Yes (.844)	Yes (.232)	

These results are consistent with our hypothesis, showing that, independent of gender, H* F_0 peak positions in late target words were later in their vowels when not followed by a boundary tone (*no-bnd-late-daily*), compared to vowels followed by a boundary tone (*bnd*). These results hold for both target words (*dagada* and *dada*), with the exception of *dada* by male speakers, where statistical analysis was not as significant (p=0.009), further supporting the hypothesis that the location of the peak F_0 for an H* is affected more by tonal crowding [8] from boundary tones than by the serial position of the word, or the number of syllables in the accented word.

Although this general pattern of greater H* F_0 peak delay in non-boundary conditions holds across both genders, there were some gender differences. For example, as noted above, differences between *no-bnd-late-daily* and *bnd* cases for H* F_0 peak position were less statistically significant for the target word *dada* for males. In addition, the F_0 peak position (as a percent of the vowel duration) is later for the female speakers than for the males. As expected from [3], H* F_0 peaks were more delayed than L* F_0 valleys, across target word, boundary condition and gender. Finally, there were effects on the height of the F0 peak: e.g. for male speakers the H* F_0 peak height was larger for *no-bnd-late-daily* cases compared to *bnd* cases, possibly because the immediately-following L- in the *bnd* case affected the H* peak.

4.2. *Energy*

We hypothesized that the energy change for a pitch accent might be similar across gender and pitch accent type, but different in boundary vs. non-boundary conditions, because of respiratory and subglottal pressure changes at the end of an utterance [14]. Two-way ANOVA results for the energy measure as a function of gender, pitch accent tone-type and presence of adjacent boundary tone are shown in Table 2 for the target word *dagada*, and Figure 2 shows a scatter plot for the male speakers for the H* accent. Results for *dada* were similar. Values represent the change in energy relative to the average energy for each utterance in percent, so 100% corresponds to twice the average sentence energy. For example, the first row in Table 2 shows that the average energy of the H* accented vowel for male speakers in the *no-bnd-late-daily*/*bnd* case is 115/47% higher than mean energy of the utterance, statistically significant with an effect size of 0.507. Similarly, the columns show the change in energy for a particular boundary condition given the accent type.

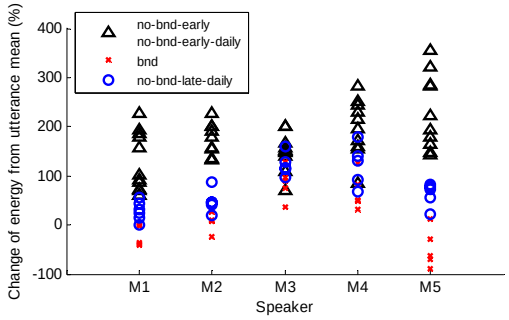

Figure 2: *Scatter plot of relative mean vowel energy of the H* accented vowel for the target word* ***dagada*** *for male speakers in three different contexts: 1. no-bnd-early/no-bnd-early-daily (triangles); 2. bnd (crosses); 3. no-bnd-late-daily (circles).*

The results were generally consistent with our hypothesis; energy values were higher for non-boundary cases than for boundary cases regardless of the type of pitch accent and independent of target word or gender. However, there were also significant differences associated with gender and pitch accent type. For example, energy values showed greater variance for males across boundary conditions, and values were higher for H*/no-acc(H-H%) than for L*/no-acc(L-L%) regardless of gender. Thus, as for F_0 peak location,

understanding the acoustic correlates of tonal targets is improved by noting the context in which the target is realized.

Table 2: *Energy of H* peaks and L* valleys for the target word* ***dagada*** *for male and female speakers. Statistical significance (s.s.) shown for no-bnd-late-daily vs. bnd and H* vs L*. Energy means and standard deviations of the target vowels are relative to the sentence mean in percent.*

ΔEnergy mean (std) in %	Males			Females		
	no-bnd-late-daily	bnd	s.s. (η^2)	no-bnd-late-dly	Bnd	s.s. (η^2)
H*	115 (62)	47 (74)	Yes (.507)	94 (60)	41 (60)	Yes (.417)
L*	55 (77)	34 (70)	No	-12 (65)	-29 (46)	No
no-acc (H-H%)	23 (54)	-15 (49)	Yes (.276)	25 (40)	24 (48)	No
no-acc (L-L%)	-52 (33)	-70 (21)	No	-46 (18)	-71 (11)	Yes (.545)
s.s. (η^2)	Yes (.784)	Yes (.746)		Yes (.789)	Yes (.879)	

4.3. *Duration*

Predictions about phrase boundary effects on duration are complex, since several different factors are at work. These include duration lengthening associated with main lexical stress [9], with the main-stress syllable of the phrase-final word [10], with the final syllable of the phrase [11] and with a pitch accent [12]. We hypothesized, among other things, that target main-stress vowels would be longer in the boundary condition than in the non-boundary condition, because of phrase-final main-stress lengthening.

Results showed that main-stress vowels were longer (by an average of 16/23 ms for males/females) in the boundary condition for *dagada*, independent of tone and gender, although for L* vowels produced by the male speakers the difference was less significant (p=0.009). The target word *dada* showed a similar pattern, although results for male speakers with H*/L* accents were less statistically significant (p=0.259/0.223). In general, female speakers showed more reliable lengthening on the main-stress syllable of the phrase-final word than did males, as well as longer duration values for the main-stress vowel of both target words independent of tonal target and boundary condition.

4.4. H_1^*-H_2^*

In [3] it was shown that for non-boundary cases, H_1^*-H_2^* values were lower for stressed vowels regardless of pitch accent, suggesting a more tense or pressed voice. In this study, ANOVA tests using speaker and tone-type as fixed factors showed similar results for the non-boundary cases except for male speakers with H* target words, where the vowel immediately following the target vowel had a lower mean H_1^*-H_2^* value than the target vowel. Data also show that for male speakers the H_1^*-H_2^* values are positively related with the F_0 values (Pearson correlation coefficient of 0.59). A similar result was found in [13] for low-pitched speakers with F_0 between 80-175 Hz.

For boundary cases, the male speakers generally showed the same relationship for the target word *dagada*, but not for *dada*, possibly because of its shorter duration. Results for female speakers seemed to be speaker-dependent, as in [3]. In the non-boundary case, female speakers on average had a lower H_1^*-H_2^* value (about 2.3 dB) on the stressed vowel, possibly indicating a "pressed" voice.

5. Conclusions

In this study of the effects of phrase boundaries on the realization of pitch accents, it was shown that the peak F_0 for H* accents occurs earlier when the target word also carries boundary tones, consistent with tonal crowding. This hypothesis is supported by the later location of the peak when the final target word is followed by *daily,* where no tonal crowding on the accented word is predicted. No such pattern was seen for L* accents, whose F_0 valley consistently occurred near the middle of the accented syllable in all contexts. Energy was higher in the accented vowel for H* accents and generally higher for male speakers. Duration of the accented vowel was generally longer in the boundary condition, consistent with main-stress-syllable lengthening in phrase-final words, and longer for the pitch-accented condition in both the early and late the non-boundary cases, consistent with accentual lengthening; these results were more reliable for female speakers. H_1^*-H_2^* in the accented vowel for males seemed to be dependent on F_0. Overall, the results highlight the importance of taking prosodic context into account when interpreting cues to specific prosodic elements such as pitch accents.

6. Acknowledgements

This work is supported in part by NSF Grant BCS0720304 to A. Alwan, and by NSF Grant BCS0643054 to S. Shattuck-Hufnagel.

7. References

[1] J. B. Pierrehumbert, "The phonology and phonetics of English intonation", Dissertation, MIT, 1980.

[2] M. Beckman and J. B. Pierrehumbert, "Intonational structure in Japanese and English", *Phonology Yearbook 3*, 255-309.

[3] Y.-L. Shue, M. Iseli, N. Veilleux, and A. Alwan, "Pitch accent versus lexical stress: Quantifying acoustic measures related to the voice source," *Proc. Interspeech*, 2007, pp. 2625-2628.

[4] H. Kawahara, A. de Cheveigné, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high quality speech transformation: revised TEMPO in the STRAIGHT-suite," in *Proc. ICSLP*, 1998.

[5] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *JSHR*, vol. 38, pp. 121-1223, 1995.

[6] M. Iseli and A. Alwan, "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation", *Proc. ICASSP*, 2004, pp. 669-672.

[7] K. Sjölander, "Snack sound toolkit," KTH Stockholm, 2004.

[8] A. Arvaniti, D. R. Ladd, and I. Mennen, "Phonetic effects of focus and "tonal crowding" in intonation: Evidence from Greek polar questions", *Speech Communication 48*, pp. 667-696, 2006.

[9] M. E. Beckman and J. Edwards, "Articulatory evidence for differentiation stress categories", *Lab. Phon. III*, pp. 7-33, 1994.

[10] A. Turk and S. Shattuck-Hufnagel, "Phrase-final lengthening in American English", *J. Phonetics*, vol. 35(4), pp. 445-472, 2007.

[11] D. Klatt, "Structure of a phonological rule component for a synthesis-by-rule program", IEEE Trans. on Acoustics, Speech, and Signal Proc., 24, 391-398, 1976.

[12] A. E. Turk and L. White, "Structural effects on pitch accentual lengthening in English", *J. Phonetics* 27, 171-206, 2007.

[13] M. Iseli, Y.-L. Shue and A. Alwan, "Age, sex, and vowel dependencies of acoustical measures related to the voice source", *JASA*, vol. 121(4), pp. 2283-2295, April 2007.

[14] J. Slifka, "Some physiological correlates to regular and irregular phonation at the end of an utterance", *J. Voice* 20, 171-186, 2007